Towards an Advanced Entity Resolution in Data Lakes

Lamisse F. BOUABDELLI[1,2]

[1]CBI² Research laboratory, Trimane The data intelligence company, 102 Terr. Boieldieu, 92800 Puteaux, France
[2]Laboratoire LIAS - ISAE-ENSMA, Téléport 2 - 1 avenue Clément Ader BP 40109, 86961 Chasseneuil, France

Abstract

Entity Resolution (ER) is a critical challenge for maintaining data quality in data lakes, aiming to identify different descriptions that refer to the same real-world entity. We address here the problem of ER in data lakes, where their schema-less architecture and heterogeneous data sources often lead to entity duplication, inconsistency, and ambiguity, causing serious data quality issues. Although ER has been well studied both in academic research and industry, many state-of-the-art ER solutions face significant drawbacks. Existing ER solutions typically compare two entities based on attribute similarity, without taking into account that some attributes contribute more significantly than others in distinguishing entities. In addition, traditional validation methods that rely on human experts are often error-prone, time-consuming, and costly. We propose an efficient ER approach that leverages deep learning, knowledge graph (KG), and large language model (LLM) to automate and enhance entity disambiguation. Furthermore, the matching task incorporates attribute weights, thereby improving accuracy. By integrating LLM for automated validation, this approach significantly reduces the reliance on manual expert verification while maintaining high accuracy.

Keywords

Data Lakes, Data Quality, Entity Resolution, Entity Matching, Machine Learning

1. Context and Motivation

The exponential growth in volume, velocity, and variety of data has introduced the concept of Big Data, which has significantly transformed how organizations store, process, and analyze information. To manage these large-scale heterogeneous datasets, organizations have adopted data lakes, scalable storage systems designed to ingest structured, semi-structured, and unstructured data in its raw format without requiring a predefined schema. This schema-less architecture offers flexibility and scalability, making data lakes an attractive solution for organizations seeking to integrate data from different sources[1, 2, 3].

Today, organizations rely on their data for strategic decision making, utilizing advanced analytics, machine learning, and business intelligence (BI) tools to gain operational efficiency and competitive advantage. Typically, data from multiple sources are stored in a data lake, then cleaned and transformed before being ingested into a data warehouse, where they are used for analytics and decision making as illustrated in Figure 1.

However, Datasets originating from heterogeneous sources inevitably introduce data quality issues, as they often differ in structure, format, schema and semantics, leading to inconsistencies, duplicate records, missing attributes, and lack of standardization[4]. Such issues degrade the accuracy and reliability of analytical outputs, resulting incorrect decision making.

One of the most critical challenges in data lakes is entity ambiguity, which occurs when multiple records from different datasets and sources represent the same real-world entity but appear in different formats. Conversely, highly similar records may correspond to different entities. Figure 2 illustrates these scenarios:

- Case (a) shows two records from distinct sources that refer to the same person

ER2025: Companion Proceedings of the 44th International Conference on Conceptual Modeling: Industrial Track, ER Forum, 8th SCME, Doctoral Consortium, Tutorials, Project Exhibitions, Posters and Demos, October 20-23, 2025, Poitiers, France

lamisse.bouabdelli@trimane.fr (L. F. BOUABDELLI)
0009-0002-9010-3128 (L. F. BOUABDELLI)

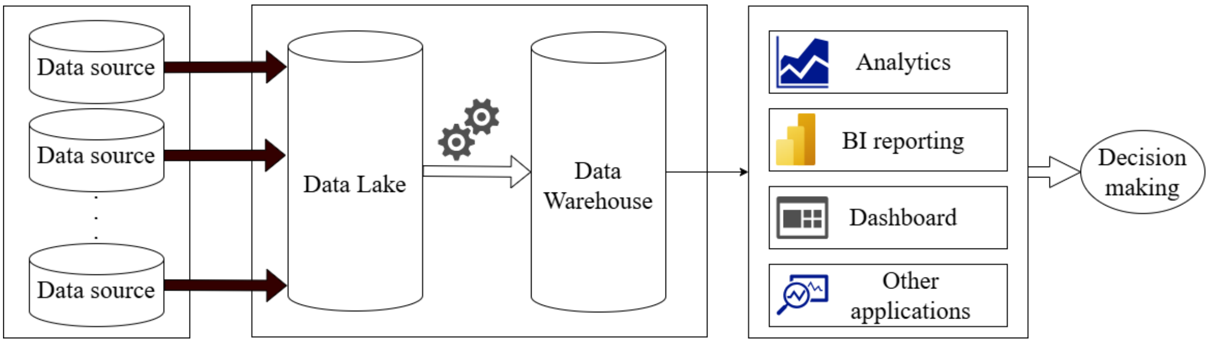
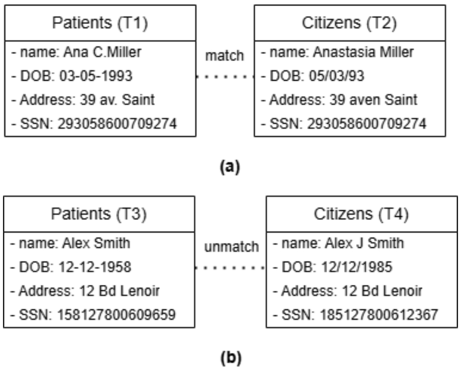


Figure 1: Entity resolution within the data management architecture


Patients (T1)
- name: Ana C.Miller
- DOB: 03-05-1993
- Address: 39 av. Saint
- SSN: 293058600709274

match

Citizens (T2)
- name: Anastasia Miller
- DOB: 05/03/93
- Address: 39 aven Saint
- SSN: 293058600709274

(a)

Patients (T3)
- name: Alex Smith
- DOB: 12-12-1958
- Address: 12 Bd Lenoir
- SSN: 158127800609659

unmatch

Citizens (T4)
- name: Alex J Smith
- DOB: 12/12/1985
- Address: 12 Bd Lenoir
- SSN: 185127800612367

(b)

Figure 2: Entity resolution examples: (a) records representing the same person, (b) different individuals despite similar attributes.

- Case (b) shows two records from distinct sources with quite similar names, identical addresses, and nearly identical birth dates, but in fact they represent two different individuals. Mismatching them would lead to false positive

In order to improve these issues, organizations require a robust Entity resolution (ER) approach. ER [5] is the process of determining whether two entities refer to the same real-world entity [6, 7]. The term *entity* refers to a distinct and identifiable unit that represents an object, a person, a place, or a concept of the real world. An entity has attributes that describe its characteristics. The term *resolution* is used because ER is fundamentally a decision-making process to resolve the question: Do the descriptions refer to the same or different entities? [8]. ER is also defined as "the process of identifying records judged to represent the same real-world entity" [9].

Since organizations rely on data for important decision-making, the need for a robust ER solution has never been more critical. In high-stakes domains such as healthcare, finance, and e-commerce, errors in ER can have severe consequences, from incorrect patient records leading to misdiagnoses, to fraudulent financial transactions, or misattribute customer data affecting business decisions.

2. Related Work

Entity resolution has been a key area of interest both in academic research [5, 7] and industry, evolving significantly from traditional similarity measures to machine learning techniques that have shown an improvement in matching performance [6]. By the late 2010s, deep learning became a key area of research in data matching [10, 11, 12]. Other research has studied ER using graph-based methods [13], and more recently experimented with LLM [10].

The industry has proposed many ER solutions using machine learning and artificial intelligence. Among these solutions, Senzing [14], designed for entity matching, combines ML clustering and AI. Quantexa

[15] employs ML and AI techniques, offers entity linkage; however, its reliance on complex graph structures poses implementation challenges for non-expert users. AWS Glue [16], a cloud-based ER solution, integrates entity resolution within broader ETL workflows. Its scalability and seamless integration with AWS services make it a powerful tool. DataWalk [17], on the other hand, is a unified graph and AI platform for data management, analysis, and investigative intelligence, which includes entity resolution software.

Despite the advancements offered by these tools, industrial ER solutions still face key limitations. A critical drawback is the lack of attribute weighting, where all entity attributes are treated equally despite varying levels of significance, which can lead to suboptimal matching results. Furthermore, the validation phase often relies on manual intervention, thereby increasing operational costs and time. This dependence not only reduces the efficiency of these solutions, but also introduces human error. Due to these issues, there is a clear need for improvement in industrial ER tools to better address these challenges. Enhanced attribute weighting mechanisms and the use of reliable automated validation could significantly refine the accuracy and efficiency of entity resolution in industry.

ER has been a main focus in research for decades and is still receiving attention [6]. It started with domain experts matching entities by hand [18]. Now, with advances in technology, machine learning-based approaches have been introduced, using supervised and unsupervised learning techniques to improve ER [6]. Methods such as Support Vector Machines [19, 20] classify entity pairs based on engineered similarity features, while Random Forests [21] employ ensemble learning to improve classification performance. However, these models require extensive feature engineering and struggle with unseen entity variations, limiting their adaptability to large and evolving datasets. Transformer and pre-trained models like BERT [22] and RoBERTa [23] revolutionized natural language processing. Studies have explored entity matching using pre-trained models[24, 25]. More recently, deep learning models have significantly advanced entity resolution by capturing contextual dependencies between entity attributes. DeepMatcher [11] applies bidirectional LSTM [26] with attention mechanisms to learn entity similarity from labeled data, while Ditto [12] uses transformer-based architectures to fine-tune pre-trained models on ER tasks. Ditto brings some optimizations that require domain knowledge. These deep learning-based methods are based on text sequences for matching. They use different methods for attribute embedding and attribute similarity representation. Furthermore, HierGAT (Hierarchical Graph Attention Networks) [13] enhances entity matching by incorporating graph-based relationships, demonstrating the potential of graph neural networks (GNNs) for ER problems. Despite their improvements in precision and recall, deep learning-based methods often overlook the importance of attribute weighting and struggle with explainability, posing challenges for real-world adoption. The advent of LLM such as Llama and GPT has further pushed the boundaries of ER by enabling zero-shot and few-shot learning for entity matching tasks [10]. Although LLM have shown strong performance, their effectiveness remains highly dependent on domain-specific fine-tuning and prompt engineering, making them computationally expensive and less adaptable to structured relational datasets. Moreover, existing LLM-based approaches do not inherently model inter-entity relationships, which limits their applicability in graph-based ER scenarios.

To clarify, an LLM that performs matching based solely on textual attributes might miss the underlying relationships between entities. For instance, consider a father and son who share the same last name and home address. An LLM could mistakenly classify them as the same person due to the high textual similarity of their attributes. However, the crucial relationship (father–son) indicates they are related but distinct individuals. This relational nuance cannot be captured by the LLM alone. In contrast, a knowledge graph can explicitly represent such relationships, enabling the system to recognize them as separate entities. This example demonstrates why relying exclusively on LLM can be problematic in graph-based ER settings: LLM lack explicit, structured mechanisms to represent and reason over inter-entity relationships.

In contrast, research efforts have also explored rule-based methods [27] that require designing rules and setting thresholds and crowd-sourcing-based ER methods[28], which require extensive manual

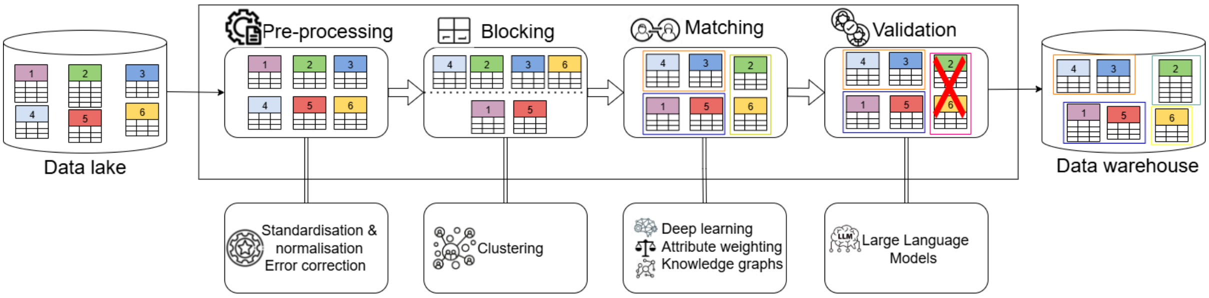


Figure 3: Pipeline of the proposed approach

intervention or rely on human annotators to validate entity matches. As ER continues to evolve, our research focus on hybrid approaches that combine deep learning, knowledge graphs, and pre-trained LLM, leveraging the strengths of each paradigm to improve entity resolution across diverse real-world datasets.

3. Research Questions

The growing adoption of data lake within organizations highlights a critical research challenge: how to identify records originating from different sources and datasets that correspond to the same real-world entity. While existing ER approaches while effective in controlled academic settings, fall short in enterprise and data lake contexts due to several limitations since they do not take into account the weight of the attributes, leading to mismatched possibility. In addition, they rely on manual validation, making the process expensive and time consuming.

These limitations raise the following research questions:

- RQ1: How can we design a robust ER pipeline tailored for a data lake involving heterogeneous sources?
- RQ2: How can we effectively incorporate attribute weighting to improve the performance of entity matching?
- RQ3: How can the ER process be designed to minimize manual intervention?

4. Research Plan and Methodology

This section introduces our proposed entity resolution approach. Figure 3 illustrates the pipeline of our proposed method. This pipeline is designed to operate within the architectural framework shown in Figure 1, especially between the data lake and the data warehouse.

Our approach is inspired by established techniques in the literature [7], but introduces key adaptations to improve the entity resolution process. The process consists of four main steps: 1) Pre-processing ensures data quality by standardizing formats, correcting mistyping errors, handling missing values, and normalizing variations. 2) Blocking aims to reduce computational complexity by grouping similar records in the same block to limit entity comparisons to subsets. 3) Matching compares records within the same block in order to identify records that correspond to the same real-world entity. 4) Validation is traditionally carried out by domain experts, which is often time-consuming and costly. To address this, we propose an automated validation mechanism that significantly reduces manual effort.

The following subsections provide a detailed explanation of each phase of the pipeline, including the specific techniques and methodology used.

4.1. Pre-processing

Pre-processing is a critical step to ensure quality of data, which is essential for entity resolution, but our contribution does not lie in this phase. In our approach we will adopt state-of-the-art techniques widely used in the literature [7, 29] such as standardization, where data formats such as dates and addresses are unified, correcting mistyping errors, and identifying missing values. Additionally, linguistic normalization is applied to unify abbreviations, acronyms, and variations of entity names, plus special character removal and the elimination of unnecessary punctuations, symbols, and whitespace.

Given that real-world data are often noisy and incomplete, we aim to improve the data quality by assuring accuracy (ensuring that data correctly reflect real-world entities), consistency (ensuring that data are harmonized and uniformed across multiple sources), correctness (verify data validity), and completeness (assessing whether all the essential information is present). Completeness is further categorized into: total completeness means no missing data, partial completeness some missing data, but it will not affect the processes and the information remains exploitable, critical completeness where essential data are missing.

The goal of pre-processing is to enhance data quality for the next steps. The output of the pre-processing is clean data, for the purpose of reducing the number of sets for the matching, for this reason we introduce our next phase.

4.2. Blocking

Blocking is an optimization step designed to reduce the number of comparisons between entity pairs, thus significantly reducing computational costs. Instead of evaluating all possible entity pairs, blocking groups similar records together in the same block. This ensures that only the most relevant subsets are considered for a detailed matching [6].

Various blocking techniques have been explored in the literature [6, 30, 24], each with its own advantages. In our approach, we plan to investigate clustering-based blocking approaches already used in literature in order to group similar records based on their attributes.

After grouping potentially similar records in blocks and reducing computational complexity as a result, limiting entity comparisons to subsets that are ready for the next step.

4.3. Matching

The matching phase constitutes the most critical step in entity resolution, as it seeks to identify records that correspond to the same real-world entity, despite variations in their descriptions, a phenomenon known as synonymy, as illustrated in Figure 2(a). In contrast, it is equally crucial to differentiate records that may exhibit similar attributes but actually represent distinct entities, a challenge called homonymy or entity collision, as shown in Figure 2(b).

Our proposed matching approach incorporates attribute weighting, recognizing that certain attributes contribute more significantly than others in distinguishing entities. We acknowledge that previous ER approaches have incorporated attribute importance through weighted similarity. Notably, recent graph-based models like [13] employ attention mechanisms to identify the most discriminative attributes. To determine these weights, we investigate two methods: (i) human-based, where domain experts assign weights for attributes based on their discriminative power, and (ii) machine learning techniques that automatically infer weights from labeled or partially labeled training data.

To further enhance matching accuracy, our approach combines deep learning techniques using pre-trained language models such as BERT. The similarity between two attribute values is then calculated using cosine similarity (or alternative distance metrics if appropriate). We plan to combine weighting mechanisms with deep learning techniques and knowledge graphs that capture the relationships between entities that are likely to match. This hybrid approach will ultimately improve ER by ensuring a more context-aware, semantically enriched, and structurally informed matching process, leading to higher precision and reduced false positives.

4.3.1. Problem Formulation

Figure 4 illustrates a scenario in which existing entity resolution solutions may incorrectly merge two distinct entities due to high similarity in certain attributes. Specifically, Entity 1 (T3) in the patient table

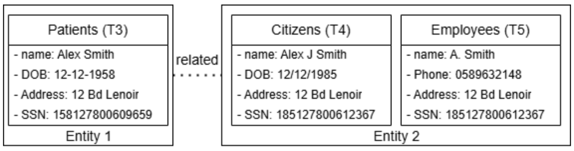


Figure 4: Example of entity resolution

and Entity 2 (T4 in the citizen table and T5 in the employee table) share common data points, such as address and name variations, making them appear as potential duplicates. The question arises: *How does our proposed matching approach differ and why is it more effective?*

To formalize the problem, we define the following.

- A dataset consisting of multiple tables T, where each table contains a set of attributes denoted as:

$$A = \{a_1, a_2, \ldots, a_n\}. \tag{1}$$

- Attributes, such as in our example name, date of birth, address, and Social Security Number (SSN).
- A weight function $w : A \rightarrow [0, 1]$ that assigns a weight to each attribute based on its discriminative power. For instance, attributes like SSN have a high weight due to their uniqueness:

$$\begin{aligned} w(\text{SSN}) &= 0.9, \\ w(\text{name}) &= 0.7, \\ w(\text{DOB}) &= 0.5, \\ w(\text{address}) &= 0.4. \end{aligned} \tag{2}$$

Matching Computation: Given two records T_i and T_j, we calculate a similarity score for each attribute using deep learning techniques to measure the degree of correspondence between attribute values. This results in a similarity vector:

$$\text{match}(T_3, T_4) = \{s_1, s_2, \ldots, s_n\}. \tag{3}$$

For example:

$$\text{match}(T_3, T_4) = \{0.6, 0.8, 1.0, 0.4\}. \tag{4}$$

where each s_k represents the similarity score for the attribute a_k.

To compute the final matching score, we apply a weighted sum:

$$S(T_3, T_4) = \frac{1}{n} \sum_{k=1}^{n} w(a_k) \cdot s_k. \tag{5}$$

Alternatively, beyond a simple summation, we plan to investigate the use of the Skyline operator [31], which is considered an optimization solution that selects non-dominated matches based on Pareto optimality.

By incorporating attribute weighting and deep learning-based similarity computation. We aim that our approach will significantly reduce false positives while improving the precision of entity resolution.

4.3.2. Capturing Relationships with Knowledge Graphs

Figure 4 illustrates an example in which existing entity resolution methods struggle, often erroneously matching two entities that are, in fact, distinct. However, our method goes beyond similarity matching by using knowledge graphs to detect relationships between entities rather than incorrectly merging them.

By integrating knowledge graphs, our approach captures semantic relationships between entities. In this example, instead of falsely concluding that Entity 1 and Entity 2 are the same person, we find a related relationships, they are likely father and son. Social Security Numbers (SSN) differ, but name, address, and other attributes share similarities, which can mislead conventional entity resolution methods.

To model this, we define $E = \{e_1, e_2, \ldots, e_n\}$ as the set of entities. R as the set of possible relationships between entities, where each relationship is defined as a directed edge $r(e_i, e_j)$ in the knowledge graph. A similarity function $S(T_i, T_j)$ that computes the weighted similarity between records, capturing both direct attribute matches and inferred relationships.

Using this structure, our method assigns relationship probabilities instead of merely merging entities. The system recognizes that while Entity 1 and Entity 2 are distinct, they are related, thus preventing false positives in entity resolution.

We believe that by combining deep learning techniques for attribute matching with knowledge graphs for relationship inference, our approach will achieve higher accuracy in distinguishing similar but distinct entities while assuring the preservation of important relationships rather than erroneous resolved entities. Plus scalability in handling complexities in real-world data.

Lastly, after finding entities that match and for the purpose of affirming if the resolved entities are correctly matched, we present our next step for entity validation.

4.4. Validation

The validation phase in our pipeline aims to verify that the resolved entities match correctly. Traditionally, this step relies on domain experts to manually verify. This approach, while reliable, is highly time-consuming, costly, and prone to human errors, especially when dealing with large-scale datasets.

To overcome these limitations, we propose an automated validation mechanism using LLM. In order to check uncertain cases which did not exceed a predefined similarity threshold in the matching phase without human intervention. Our approach utilizes the reasoning and contextual understanding capabilities of LLM, allowing them to provide a final layer of confidence in the entity resolution process.

We are aware that LLM can be used during the matching phase. However, we deliberately restrict their use to the validation phase because of considerations of cost, scalability, and explainability. Running an LLM on every candidate pair during matching would be computationally expensive and inefficient, especially when processing millions of comparisons. In contrast, using LLM only on a reduced subset of record pairs, those that survived earlier blocking and matching a better balance between accuracy and performance. This approach allows us to benefit from LLM sophisticated reasoning capabilities. LLM-based validation step provides a final layer of confidence by validating only the top-ranked candidate pairs.

By automating validation, we eliminate the reliance on domain experts for this task, consequently reducing human effort and operational costs. The approach is highly scalable, capable of efficiently validating millions of records, which makes it well suited for large-scale entity resolution. Using the contextual reasoning capabilities of LLM, our method aims to ensure high accuracy by minimizing false positives and false negatives. Furthermore, the execution speed of our validation mechanism is faster than manual methods, enabling real-time or near-real-time verification.

In summary, by integrating deep learning, knowledge graphs, and LLM, our entity resolution approach aims to ensure a more efficient, scalable, and reliable validation process.

5. Next Research Step and Expected Final Contribution

Data quality is a critical challenge in data lakes. Therefore, entity resolution is crucial to enhance data quality which is essential for making optimal decisions.

In this paper, we propose a novel entity resolution approach designed to improve data quality, scalability, and automation in data lakes. Our solution uses deep learning, to improve entity matching, knowledge graphs, to capture relationships between entities and LLM to reduce human intervention in the validation phase.

Our approach presents a potentially effective improvement to existing entity resolution solutions, but its true performance and efficiency can only be validated through real-world implementation and experimentation.

Since our work is currently a theoretical proposition, our next step is to implement this approach and conduct a comprehensive evaluation against existing solutions. We aim to demonstrate its effectiveness in the real-world and ultimately contribute to the advancement of entity resolution.

At this stage, we assume that data sources share an aligned schema, allowing attributes to be directly compared across sources. In future work, we aim to relax this assumption by explicitly addressing schema heterogeneity.

While our approach focuses on the identification of duplicate entities, we acknowledge that the subsequent step data fusion (merging duplicate records into unified representations) is not addressed in this paper. Data fusion is a critical and non trivial component of the ER pipeline, and we plan to investigate scalable and context-aware fusion strategies as part of future work.

However, we note that data fusion has already been explored in previous research efforts [32, 33, 34], where our team explored merging duplicate records in data lakes using ontology-driven integration. Building upon such foundations, our future efforts will aim to incorporate a robust, semantically informed fusion module to complete the ER pipeline.

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Fatma Abdelhedi, Allel Hadjali, and Slimane Hammoudi, for their guidance and support. Special thanks to Gilles Zurfluh for his valuable advice and mentorship.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT-5 in order to: Grammar and spelling check.

References

[1] P. N. Sawadogo, E. Scholly, C. Favre, E. Ferey, S. Loudcher, J. Darmont, Metadata systems for data lakes: Models and features, in: New Trends in Databases and Information Systems: ADBIS 2019 Short Papers, Workshops BBIGAP, QAUCA, SemBDM, SIMPDA, M2P, MADEISD, and Doctoral Consortium, Bled, Slovenia, September 8–11, 2019, Proceedings 23, Springer, 2019, pp. 440–451.

[2] F. Ravat, Y. Zhao, Data lakes: Trends and perspectives, in: International Conference on Database and Expert Systems Applications, Springer, 2019, pp. 304–313.

[3] P. Sawadogo, J. Darmont, On data lake architectures and metadata management, Journal of Intelligent Information Systems 56 (2021) 97–120.

[4] M. H. Moslemi, A. Mousavi, B. Behkamal, M. Milani, Heterogeneity in entity matching: A survey and experimental analysis, arXiv preprint arXiv:2508.08076 (2025).

[5] N. Barlaug, J. A. Gulla, Neural networks for entity matching: A survey, ACM Transactions on Knowledge Discovery from Data (TKDD) 15 (2021) 1–37. doi:10.1145/3442200.

[6] V. Christophides, V. Efthymiou, T. Palpanas, G. Papadakis, K. Stefanidis, An overview of end-to-end entity resolution for big data, ACM Computing Surveys (CSUR) 53 (2020) 1–42. doi:10.1145/3418896.

[7] P. Christen, Data Matching, Springer: Data-centric systems and applications, 2012. doi:10.1007/978-3-642-31164-2.

[8] J. Talburt, Entity Resolution and Information Quality, Elsevier, 2011. doi:10.1016/C2009-0-63396-1.

[9] O. Benjelloun, H. Garcia-Molina, H. Gong, H. Kawai, T. E. Larson, D. Menestrina, S. Thavisomboon, D-swoosh: A family of algorithms for generic, distributed entity resolution, in: 27th International Conference on Distributed Computing Systems (ICDCS'07), IEEE, 2007, pp. 37–37. doi:10.1109/ICDCS.2007.96.

[10] R. Peeters, A. Steiner, C. Bizer, Entity matching using large language models, arXiv preprint arXiv:2310.11244 (2024).

[11] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, V. Raghavendra, Deep learning for entity matching: A design space exploration, in: Proceedings of the 2018 international conference on management of data, 2018, pp. 19–34. doi:10.1145/3183713.3196926.

[12] Y. Li, J. Li, Y. Suhara, A. Doan, W.-C. Tan, Deep entity matching with pre-trained language models, Proceedings of the VLDB Endowment 14 (2020) 50–60. doi:10.14778/3421424.3421431.

[13] D. Yao, Y. Gu, G. Cong, H. Jin, X. Lv, Entity resolution with hierarchical graph attention networks, in: Proceedings of the 2022 International Conference on Management of Data, 2022, pp. 429–442. doi:10.1145/3514221.3517872.

[14] I. Senzing, Senzing – entity resolution software, https://senzing.com/, 2025.

[15] Quantexa, Quantexa, https://www.quantexa.com/fr/, 2025.

[16] AWS, Aws glue, https://aws.amazon.com/fr/glue/, 2017.

[17] Datawalk, Data walk entity resolution, https://datawalk.com/solutions/entity-resolution/, 2025.

[18] I. P. Fellegi, A. B. Sunter, A theory for record linkage, Journal of the American statistical association 64 (1969) 1183–1210.

[19] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (1995) 273–297. doi:10.1007/BF00994018.

[20] M. Bilenko, R. Mooney, Adaptive duplicate detection using learnable string similarity measures, in: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, 2003, pp. 39–48. doi:10.1145/956750.956759.

[21] L. Breiman, Random forests, Machine learning 45 (2001) 5–32. doi:10.1023/A:1010933404324.

[22] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018) 15. doi:10.48550/arXiv.1810.04805.

[23] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019). doi:10.48550/arXiv.1907.11692.

[24] M. Paganelli, D. Tiano, F. Guerra, A multi-facet analysis of bert-based entity matching models, The VLDB Journal 33 (2024) 1039–1064.

[25] Y. Li, J. Li, Y. Suhara, J. Wang, W. Hirota, W.-c. Tan, Deep entity matching: Challenges and opportunities, Journal of Data and Information Quality (JDIQ) 13 (2021) 1–17. doi:10.1145/3431816.

[26] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[27] R. Singh, V. Meduri, A. Elmagarmid, S. Madden, P. Papotti, J.-A. Quiané-Ruiz, A. Solar-Lezama, N. Tang, Generating concise entity matching rules, in: Proceedings of the 2017 ACM International Conference on Management of Data, 2017, pp. 1635–1638. doi:10.1145/3035918.3058739.

[28] J. Wang, T. Kraska, M. J. Franklin, J. Feng, Crowder: Crowdsourcing entity resolution, arXiv preprint arXiv:1208.1927 (2012).

[29] S. Vijayarani, M. J. Ilamathi, M. Nithya, et al., Preprocessing techniques for text mining-an overview, International Journal of Computer Science & Communication Networks 5 (2015) 7–16.

[30] D. Skoutas, E. Thanos, T. Palpanas, A survey of blocking and filtering techniques for entity resolution, arXiv preprint arXiv:1905.06167 (2019). doi:10.48550/arXiv.1905.06167.

[31] S. Borzsony, D. Kossmann, K. Stocker, The skyline operator, in: Proceedings 17th international conference on data engineering, IEEE, 2001, pp. 421 – 430. doi:10.1109/ICDE.2001.914855.

[32] F. Abdelhedi, R. Jemmali, G. Zurfluh, Data Ingestion from a Data Lake: The Case of Document-oriented NoSQL Databases, in: J. Filipe, M. Smialek, A. Brodsky, S. Hammoudi (Eds.), Proceedings of the 24th International Conference on Enterprise Information Systems - ICEIS 2022 ; ISBN 978-989-758-569-2 ; ISSN 2184-4992, volume 1: ICEIS, SCITEPRESS : Science and Technology Publications, Online Streaming, France, 2022, pp. 226–233. URL: https://hal.science/hal-03758340. doi:10.5220/0011068300003179.

[33] F. Abdelhedi, R. Jemmali, G. Zurfluh, DLToDW: Transferring Relational and NoSQL Databases from a Data Lake, SN Computer Science 3 (2022) article 381. URL: https://hal.science/hal-03758354. doi:10.1007/s42979-022-01287-7.

[34] F. Abdelhedi, R. Jemmali, G. Zurfluh, Ingestion of a data lake into a nosql data warehouse: The case of relational databases., in: KMIS, 2021, pp. 64–72.