Detailed Review and Analysis of

"A Review on Data Normalization Techniques"

Sanjana Bishwokarma, Patan College for Professional Studies

February 6, 2026

Abstract: *The exponential growth of data from multiple heterogeneous sources has led to the generation of massive datasets, often containing redundant or duplicate records with slight variations. Such inconsistencies not only reduce data quality but also hinder accurate analysis and decision-making processes. Data normalization specifically, the process of identifying and merging similar or duplicate records into a single, representative entity has emerged as a critical step in ensuring data integrity and enhancing its utility for mining, analytics, and knowledge discovery. This paper presents a comprehensive review of existing data normalization techniques, examining their methodologies, advantages, and limitations in handling diverse and complex datasets.*

1. INTRODUCTION

The introduction sets the stage for web data integration, where massive amounts of structured data (e.g., from web databases, tables, or warehouses like Google Scholar or Bing Shopping) are collected. Key challenges include:

- **Redundancy**: Search queries often return duplicate records from different sources, making analysis cumbersome.
- **Normalization Importance**: Essential for domains like research publications (e.g., CiteSeer, Google Scholar), where data must be de-duplicated and presented clearly. Desirable outcomes include "best match search" and de-duplication to avoid user frustration from ad-hoc or incomplete results.
- **Challenges**: Data conflicts arise from errors, incompleteness, varying formats, or missing attributes. For instance

- author names might be abbreviated differently, venues shortened, or pages omitted.

2. LITERATURE SURVREY

This section critically reviews prior work (2001–2019), identifying gaps that the proposed system addresses. Key papers include:

- **Culotta et al. (2007)**: First to propose "canonicalization" (normalization) using string edit distance for central records, parameter optimization, and feature-based improvements. Limitation: No value-level normalization, leading to repetitive data.
- **Swoosh (Benjelloun et al., 2009)**: Frames deduplication as entity resolution with match-merge functions. Efficient for de-duplication but increases complexity and doesn't produce fully normalized records.
- **Wick et al. (2008)**: Integrates schema matching, coreference resolution, and canonicalization via a discriminatively-trained model. Limitation: High complexity; field-level only.
- **Tejada et al. (2001)**: "Object normalization" from web sources using attribute and string ranking with user confidence scores.
- **Wang et al. (2013)**: Hybrid framework for shopping data normalization, handling missing values and corrections at field level (global, not value-specific).
- **Chaturvedi et al. (2013)**: Focuses on pattern discovery in duplicates rather than normalization, but applicable for standardization.
- **Dragut et al. (2006)**: Automatic label normalization for query interfaces (field-level labeling).
- **Raunich et al. (2011)**: ATOM system for ontology merging (record normalization), but requires user involvement—authors advocate for more automation.
- **Dong et al. (2019)**: Multi-level (record, field, value) normalization using string operations. Limitation: No NLP; doesn't link related entities for richer representations.

Common limitations: Over-reliance on string functions, lack of NLP, user dependency, and incomplete value-level handling.

3. PROBLEM FPRMULATION

Formally defines the problem: For a real-world entity $E1$with duplicate records $Re = \{R1, R2, \ldots, Rp\}$, each with fields $FS = \{f1, f2, \ldots, fq\}$and values $ri[fi]$, generate a descriptive normalized record using NLP. Additionally, link similar entities via field/value matching for fused, informative data.

4. PROPOSED METHODOLOGY

The core contribution: A system that processes redundant records into normalized, linked ones via five stages.

- **Preliminaries**:
    - **Frequency Ranker (FR)**: Ranks units by occurrence: $FR(U) = [u1, u2, \ldots, up]$(descending frequency).
    - **Length Ranker (LR)**: Ranks by character count: $LR(U) = [u1, u2, \ldots, up]$(descending length).
    - **Centroid Ranker**: Computes unit centrality: $UCS(u) = \frac{1}{|U'|^2} \sum_{v \in U'} Au \cdot Av \cdot SM(u, v)$, where $U'$is distinct units, Au/Avfrequencies, and $SM$is similarity (e.g., edit-distance or bigram).
    - **Similarity Measures**:
        - Edit-distance: $Sim - ed(a, b) = \frac{ed(a,b)}{\max(|a|,|b|)}$.
        - Bigram: $Sim - bigram(a, b) = \frac{2 \cdot |bigram(a) \cap bigram(b)|}{|bigram(a)| + |bigram(b)|}$.
    - **Feature-Based Rankers**: Strategy (binary indicators for ranking criteria) and Text (checks acronyms/abbreviations, e.g., "conf" for "conference", "VLDB" for "Very Large Databases").
    - **Collocations**: Sequences with low IDF; sub-collocations (substrings); template collocations (high IDF); twin templates (paired with probability thresholds).

- **System Architecture**: Input: Redundant records. Output: Normalized, linked records. Stages:

1. **Data Preprocessing**: Extract records and fields (e.g., from citation: author, title, venue, date, pages).

2. **Record-Level Normalization**: Select complete records (all fields present) for filtering.

3. **Field-Level Normalization**: Combine descriptive field features using rankers.

4. **Value-Level Normalization**: Extract values, replace abbreviations via Mining Abbreviation-Definition Pairs; identify collocations via Mining Template Collocation-Sub Collocation Pairs (MTS); use NLP n-grams for nouns and aberrations.

5. **Field-Based Clusters**: Link records by normalized field values.

- **System Description** (Detailed Example): Using the Halevy et al. citation, preprocess to separate fields, select complete records, normalize fields (e.g., venue via rankers), values (e.g., expand "VLDB"), and cluster similar entities.

5. CONCLUSION

Existing systems normalize via de-duplication at varying levels but often require user input and lack NLP. The proposed system automates multi-level normalization (record → field → value), increasing precision, and adds linking for representative data. It uses NLP n-grams for value accuracy with reduced processing. Future work could minimize user involvement further.

REFERENCES

[1] Y. Dong, E. C. Dragut, and W. Meng, "Normalization of Duplicate Records from Multiple Sources," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 4, pp. 769–782, Apr. 2019. DOI: 10.1109/TKDE.2018.2844176 (Available at: https://ieeexplore.ieee.org/document/8372637 or open PDF versions on ResearchGate/Temple CIS site)

[2] A. Culotta, M. Wick, R. Hall, M. Marzilli, and A. McCallum, "Canonicalization of database records using adaptive similarity measures," in *Proc. 13th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD '07)*, San Jose, CA, USA, 2007, pp. 201–209. (PDF available at: https://people.cs.umass.edu/~mccallum/papers/canonical-kdd07.pdf)

[3] O. Benjelloun, H. Garcia-Molina, D. Menestrina , Q. Su, S. E. Whang, and J. Widom, "Swoosh: A generic approach to entity resolution," *The VLDB Journal*, vol. 18, no. 1, pp.

255–276, Feb. 2009. DOI: 10.1007/s00778-008-0098-x (Original preprint/PDF: http://infolab.stanford.edu/serf/swoosh_vldbj.pdf)

[4] M. L. Wick, K. Rohanimanesh, K. Schultz, and A. McCallum, "A unified approach for schema matching, coreference and canonicalization," in *Proc. 14th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD '08)*, Las Vegas, NV, USA, 2008, pp. 722–730.

[5] S. Tejada, C. A. Knoblock, and S. Minton, "Learning object identification rules for information integration," *Information Systems*, vol. 26, no. 8, pp. 607–633, Dec. 2001. DOI: 10.1016/S0306-4379(01)00042-4 (Available via ScienceDirect or USC/ISI archives)

[6] L. Wang, R. Zhang, C. Sha, X. He, and A. Zhou, "A hybrid framework for product normalization in online shopping," in *Proc. 18th Int. Conf. Database Systems for Advanced Applications (DASFAA 2013)*, Part II, Wuhan, China, 2013, pp. 370–384. (Lecture Notes in Computer Science, vol. 7826)

[7] S. Chaturvedi et al., "Automating pattern discovery for rule based data standardization systems," in *Proc. 29th IEEE Int. Conf. Data Engineering (ICDE 2013)*, Brisbane, Australia, 2013, pp. 1231–1241.

[8] E. C. Dragut, C. Yu, and W. Meng, "Meaningful labeling of integrated query interfaces," in *Proc. 32nd Int. Conf. Very Large Data Bases (VLDB 2006)*, Seoul, Korea, 2006, pp. 679–690.

[9] S. Raunich and E. Rahm, "ATOM: Automatic target-driven ontology merging," in *Proc. 27th IEEE Int. Conf. Data Engineering (ICDE 2011)*, Hannover, Germany, 2011, pp. 1276–1279.