ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Special Issue , August 2019

International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

New Strategic Normalization of Duplicate Records from Multiple Sources

P. Somaraju, P Veeraswamy, K. Gopi

Assistant Professor, Department of Computer Science & Engineering, Sree Vahini Institute of Science & Technology , Tiruvuru, A.P, India

Assistant Professor, Department of Computer Science & Engineering, Sree Vahini Institute of Science & Technology , Tiruvuru, A.P, India

Assistant Professor, Department of Computer Science & Engineering, Sree Vahini Institute of Science & Technology , Tiruvuru, A.P, India

ABSTRACT: Data consolidation is a challenging issue in data integration. The usefulness of data increases when it is linked and fused with other data from numerous (Web) sources. The promise of Big Data hinges upon addressing several big data integration challenges, such as record linkage at scale, real-time data fusion, and integrating Deep Web. Although much work has been conducted on these problems, there is limited work on creating a uniform, standard record from a group of records corresponding to the same real-world entity. We refer to this task as *record normalization*. Such a record representation, coined *normalized record*, is important for both front-end and back-end applications. In this paper, we formalize the record normalization problem, present in-depth analysis of normalization granularity levels (e.g., *record*, *field*, and *value-component*) and of normalization forms (e.g., *typical* versus *complete*). We propose a comprehensive framework for computing the normalized record. The proposed framework includes a suit of record normalization methods, from naive ones, which use only the information gathered from records themselves, to complex strategies, which globally mine a group of duplicate records before selecting a value for an attribute of a normalized record. We conducted extensive empirical studies with all the proposed methods. We indicate the weaknesses and strengths of each of them and recommend

KEY WORDS: Record normalization, data quality, data fusion, web data integration, deep web, database.

I. INTRODUCTION

The Web has evolved into a data-rich repository contain- ing a large amount of structured content spread across millions of sources. The usefulness of Web data increases exponentially (e.g., building knowledge bases, Web-scale data analytics) when it is linked across numerous sources. Structured data on the Web resides in Web databases [1] and Web tables [2]. Web data integration is an important component of many applications collecting data from Web databases, such as Web data warehousing (e.g., Google and Bing Shopping; Google Scholar), data aggregation (e.g., product and service reviews), and metasearching [3]. Integration systems at Web scale need to automatically match records from different sources that refer to the same real-world entity [4], [5], [6], find the true matching records among them and turn this set of records into a standard record for the consumption of users or other applications. There is a large body of work on the *record matching problem* [7] and the *truth discovery problem* [8]. The record matching problem is also referred to as duplicate record detection [9], record linkage [10], object identification [11], entity resolution [12], or deduplication [13] and the truth discovery problem is also called as truth finding [14] or fact finding a key problem in data fusion [16], [17]. In this paper, we assume that the tasks of record matching and truth

Record normalization is important in many application domains. For example, in the research publication domain, although the integrator website, such as Citeseer or Google Scholar, contains records gathered from a variety of sources using automated extraction techniques, it must display a normalized record to users. Otherwise, it is unclear what can be presented to users: (i) present the entire group of matching records or (ii) simply present some random record from the group, to just name a couple of ad-hoc approaches. Either of these choices can lead to a frustrating experience for a

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Special Issue , August 2019

International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

user, because in (i) the user needs to sort/browse through a potentially large number of duplicate records, and in (ii) we run the risk of presenting a record with missing or incorrect pieces of data.

Thus, we can create a normalized value for venue, at the value-component level, as follows.

1) We take the value suggested previously by the field- level for venue and replace the abbreviations in it with the complete words and change it into "in proceedings 32nd international conference on Very large data bases".
2) We find that "in proceedings" is the part of the collocation "in proceedings of the".
3) We use the collocation to replace "in proceedings".
4) Finally, we get the normalized value of venue, "in proceedings of the 32nd international conference on Very large data bases".

Four records for the same publication: R_a, R_b, R_c, and R_d are extracted from different websites and R_{norm} is constructed manually.

Fields	author	title	venue	date	pages
R_a	Halevy, A.; Rajaraman A.; Ordille, J.	Data integration: the teenage years	in proc 32nd int conf on Very large data bases	2006	
R_b	A. Halevy, A. Rajaraman, J. Ordille	Data integration: the teenage years	in VLDB	2006	9-16
R_c	A. Halevy, A. Rajaraman, J. Ordille	Data integration: the teenage years	in proc 32nd conf on Very large data bases	2006	pp.9-16
R_d	A. Halevy, A. Rajaraman, J. Ordille	Data integration: the teenage years		2006	9-16
R_{norm}	Alon Halevy, Anand Rajara- man, Joann Ordille	Data integration: the teenage years	in proceedings of the 32nd international conference on Very large data bases	2006	9-16
R_{field}	A. Halevy, A. Rajaraman, J. Ordille	Data integration: the teenage years	in proc 32nd int conf on Very large data bases	2006	pp.9-16

Record Normalization Problem (RNP): Create a nor- malized record nr_e for each entity e E from the set of matching records R^e that summarizes the information about e as accurately as possible. Currently, there is not a widely accepted standard for record normalization, but there are a few prerequisites of a good normalized record:

(1) *Error-free*: A normalized record should avoid errors, such as misspellings or incorrect field values, as much as possible.
(2) *Comprehensive*: A normalized record should contain a value for each field whenever possible.
(3) *Representative*: A normalized record should reflect the commonality among the matched records.

II. SIGNIFICANCE OF THE SYSTEM

Levels of Record Normalization

We propose three levels of normalization: record, field, and value-component. Note that regardless of the chosen level of normalization, the goal is to provide users with some form of normalized record that is the easiest to grasp by an ordinary user

Record-level Normalization

The record-level normalization assumes that each record r_i R^e is a cohesive unit, in the sense that taken together the values $r_i[f_j]$ of the fields f_j in r_i give a coherent depiction of entity e. The assumption, while intuitively appealing and allows to build the theoretical underpins for constructing normalized records, needs to be taken with a grain of salt in practice. R^e contains a mixture of candidate normalized records and records with incomplete or arcane representations of e, which may be difficult to understand by ordinary users. The challenge is to select a record r_i R^e that is most likely

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 6, Special Issue , August 2019
International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

to be a reasonable candidate. The selection can be performed according to several criteria (described in Section 4.1). One elementary criterion is to demand that the selected record must have a value for each field.

III. LITERATURE SURVEY

Normalization Forms

We present two forms of normalization for a normalized record: *typical* and *complete*.

Typical Normalization

The purpose of typical normalization is to produce a nor- malized record that resemblances many of the matching records without modifying any of the field values. One way to define it is by frequency of occurrence. With this definition, the record-level normalization will yield a record representation that appears most often among the set of matching records for an entity. The field-level normalization will select the most frequent value for each field in the normalized record. Other strategies are clearly conceivable to perform typical normalization and we present additional alternatives in Section 4. The value-component level nor- malization inherently does not produce *typical* normalized records because it may create *new* values for some of the fields of the normalized records.

Complete Normalization

Complete normalization seeks to produce the normalized record with the property that the value of each of its fields is both complete (not missing component) and self- explanatory. For example, there are several different repre- sentations of an author's name, such as full name versus

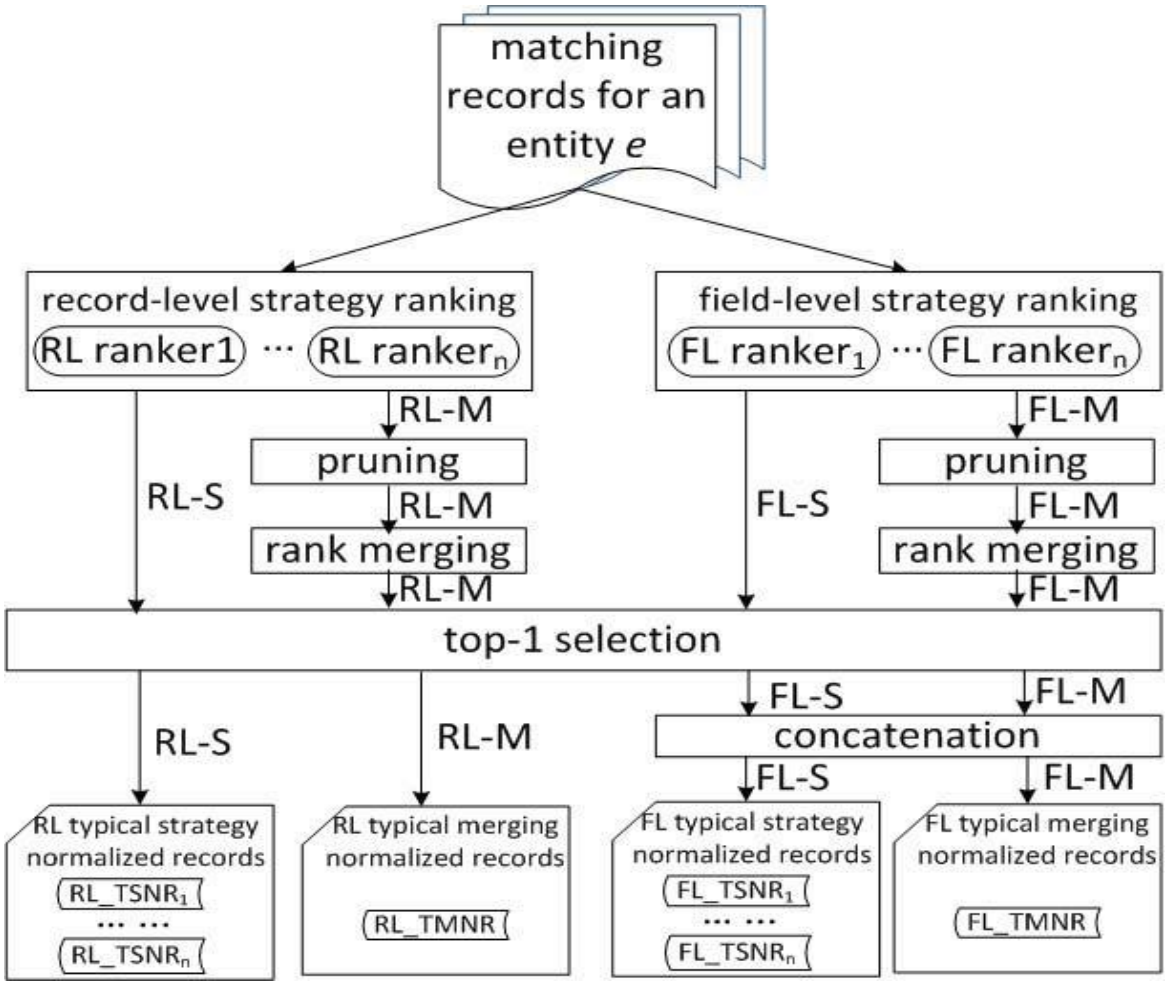


Fig. 1. The typical normalization framework.

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Special Issue , August 2019

International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

IV. METHODOLOGY

Solution Framework

We follow different steps for the two normalization forms. Fig. 1 shows the steps of the typical normalization frame- work and Fig. 2 shows those of the complete normalization framework.
In both frameworks, the input is the set of matching records R^e for an entity *e*. Different normalization strate- gies may be employed at each step in the normalization framework. Different choices will yield different normal- ized records for the same set of matching records. The normalized records are represented by parallelograms in Fig.1 and Fig.2. At every granularity level, we perform two categories of approaches: *single-strategy* and *multi-strategy* approaches. In Fig. 1 and Fig. 2, the string suffix "-S" on

Ranked List Merging

In Section, we introduced a set of single-strategy rankers each of which ranks the units (records or field values) with a different strategy. In general, a single-strategy approach does not produce satisfactory results and may even cause bias. We utilize a multi-strategy approach to combine the outcomes of several single-strategy rankers to overcome the limitations of the individual rankers. A multi-strategy approach requires an effective *rank merging algorithm* [3].

Algorithm 1 Mining Abbreviation-Definition Pairs

Input: $Val(f_j) = \{r_i[f_j] | r_i \in R^e\}$: the collection of all values of the field f_j
Output: AWP: a set of abbreviation-word pairs

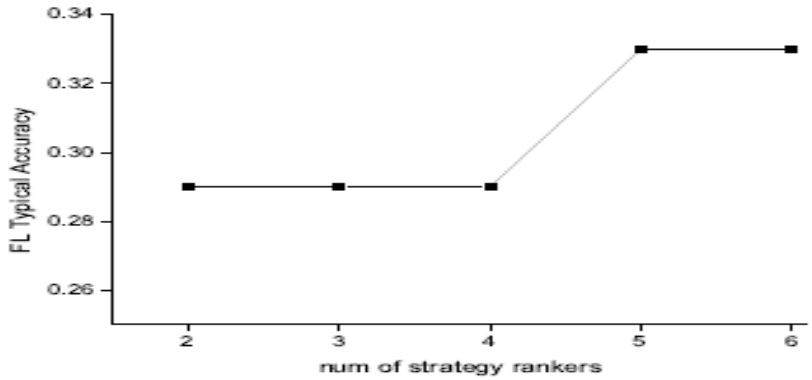
```
1: cwords = ∅; AWP = ∅;
2: pwords = tokenize(Val(f_j))
3: uwords = unique(pwords);
4: for each uword ∈ uwords do
5:    if len(uword) ≥ η_len and idf(uword, R^e) ≤ η_idf then
6:       insert uword into cwords;
7:    end if
8: end for
9: for each cword ∈ cwords do
10:   pa_words = getWordsBySameContext(
                 cword, uwords, η_pos);
11:   if pa_words ≠ ∅ then
12:      abbreviations = getAbbreviations(
                         cword, pa_words);
13:   end if
14:   if abbreviations ≠ ∅ then
15:      for each abbreviation ∈ abbreviations do
16:         insert (abbreviation, cword) into AWP;
17:      end for
18:   end if
19: end for
20: return AWP
```

V. EXPERIMENTAL RESULTS

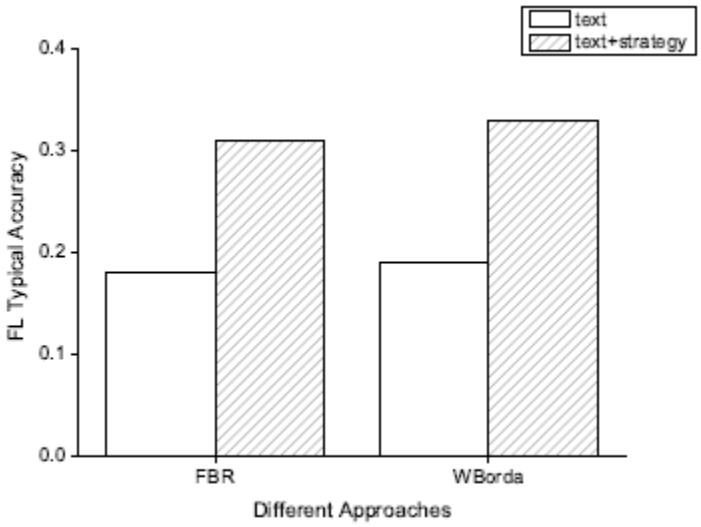
The dataset contains data about publication venue canonicalization [27]. PVCD has 3,683 publication venue values for 100 distinct real-world publication records. It is only concerned with the field venue, which is arguably the most difficult field to normal- ize, because of the presence of acronyms, abbreviations, and misspellings. We use this dataset to compare our approaches with those in [26]. The work in [26] is an instance of typical normalization, because it selects one of the duplicate records or one of the field values as the normalized record or field value, respectively. It does not attempt to create new field values or new records as normalized records. Our analysis of the dataset reveals

ISSN: 2350-0328
International Journal of Advanced Research in Science, Engineering and Technology
Vol. 6, Special Issue , August 2019
International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

that many normalized field values are labelled unreasonably. We point out some of the problems in Table 2. The column "old gold standard" shows the normalized venue values as used in the experimental study of Culotta et al. [26] and the column "new gold standard" shows them after we curates the dataset.



(a)FL Typical Accuracy Change



(a)FL Typical Normalization Accuracy Comparison

We perform experiments to evaluate the effectiveness of our approach. Summarizes the outcome of our eight approaches for the N-PVCD dataset. The first six rows in the table belong to the category of single-strategy approaches and the last two rows belong to the multi-strategy approaches. We will use the acronyms in parenthesis to refer to these approaches for the rest of this section.

The main conclusion of this experimental study is that WBorda consistently outperforms the other approaches on both FL typical normalization and VCL complete normaliza- tion. For single-strategy, FBR (Feature-based Ranker) has the best accuracy on these two forms of normalization. WBorda outperforms FBR by 6.5% on FL typical normalization and by 15.3% on VCL complete normalization. We find that the accuracy of Borda is lower than that of FBR on FL typical normalization, but higher than that of FBR on VCL complete normalization. Our explanation is that Borda treats

ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 6, Special Issue , August 2019

International Conference on Recent Advances in Science, Engineering, Technology and Management at Sree Vahini Institute of Science and Technology-Tiruvuru, Krishna Dist, A.P

uniformly the rankers and some rankers may have poor performance, which affects the final result. When rankers are assigned weights according to their contributions to the normalized record, WBorda significantly improves the normalization accuracy.

VI. CONCLUSION

In this paper, we studied the problem of record normal- ization over a set of matching records that refer to the same real-world entity. We presented three levels of nor- malization granularities (record-level, field-level and value-component level) and two forms of normalization (typi- cal normalization and complete normalization). For each form of normalization, we proposed a computational frame- work that includes both single-strategy and multi-strategy approaches. We proposed four single-strategy approaches: frequency, length, centroid, and feature-based to select the normalized record or the normalized field value.

In the future, we plan to extend our research as follows. First, conduct additional experiments using more diverse and larger datasets. The lack of appropriate datasets cur- rently has made this difficult. Second, investigate how to add an effective human-in-the-loop component into the current solution as automated solutions alone will not be able to achieve perfect accuracy. Third, develop solutions that handle numeric or more complex values.

REFERENCES

[1] K. C.-C. Chang and J. Cho, "Accessing the web: From search to integration," in *SIGMOD*, 2006, pp. 804–805.
[2] M. J. Cafarella, A. Halevy, D. Z. Wang, E. Wu, and Y. Zhang, "Webtables: Exploring the power of tables on the web," *PVLDB*, vol. 1, no. 1, pp. 538–549, 2008.
[3] W. Meng and C. Yu, *Advanced Metasearch Engine Technology*. Mor- gan & Claypool Publishers, 2010.
[4] A. Gruenheid, X. L. Dong, and D. Srivastava, "Incremental record linkage," *PVLDB*, vol. 7, no. 9, pp. 697–708, May 2014.
[5] E. K. Rezig, E. C. Dragut, M. Ouzzani, and A. K. Elmagarmid, "Query-time record linkage and fusion over web databases," in *ICDE*, 2015, pp. 42–53.
[6] W. Su, J. Wang, and F. Lochovsky, "Record matching over query results from multiple web databases," *TKDE*, vol. 22, no. 4, 2010.
[7] H. Ko¨ pcke and E. Rahm, "Frameworks for entity matching: A comparison," *DKE*, vol. 69, no. 2, pp. 197–210, 2010.
[8] X. Yin, J. Han, and S. Y. Philip, "Truth discovery with multiple conflicting information providers on the web," *ICDE*, 2008.
[9] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate record detection: A survey," *TKDE*, vol. 19, no. 1, pp. 1–16, 2007.
[10] P. Christen, "A survey of indexing techniques for scalable record linkage and deduplication," *TKDE*, vol. 24, no. 9, 2012.
[11] S. Tejada, C. A. Knoblock, and S. Minton, "Learning object identi- fication rules for information integration," *Inf. Sys.*, vol. 26, no. 8, pp. 607–633, 2001.
[12] L. Shu, A. Chen, M. Xiong, and W. Meng, "Efficient spectral neighborhood blocking for entity resolution," in *ICDE*, 2011.
[13] Y. Jiang, C. Lin, W. Meng, C. Yu, A. M. Cohen, and N. R. Smalheiser, "Rule-based deduplication of article records from bibliographic databases," *Database*, vol. 2014, 2014.
[14] X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava, "Truth finding on the deep web: Is the problem solved?" in *PVLDB*, vol. 6, no. 2, 2012, pp. 97–108.
[15] J. Pasternack and D. Roth, "Making better informed trust decisions with generalized fact-finding," in *IJCAI*, 2011, pp. 2324–2329.
[16] X. L. Dong and F. Naumann, "Data fusion: resolving data conflicts for integration," *PVLDB*, vol. 2, no. 2, pp. 1654–1655, 2009.
[17] E. K. Rezig, E. C. Dragut, M. Ouzzani, A. K. Elmagarmid, and
W. G. Aref, "ORLF: A flexible framework for online record linkage and fusion," in *ICDE*, 2016, pp. 1378–1381.
[18] X. Wang, X. L. Dong, and A. Meliou, "Data x-ray: A diagnostic tool for data errors," in *SIGMOD*, 2015, pp. 1231–1245.
[19] G. R. D. Patrick AV Hall, "Approximate string matching," *ACM Computing Surveys*, vol. 12, no. 4, pp. 381–402, 1980.
[20] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, "A comparison of string metrics for matching names and records," in *KDD workshop on data cleaning and object consolidation*, 2003, pp. 73–78.
[21] D. C. Liu and J. Nocedal, "On the limited memory bfgs method for large scale optimization," *Mathematical Programming*, vol. 45, no. 3, pp. 503–528, 1989