A Novel Normalization Technique of Identical Records from Different Resources

Treesa Jeemol

Student, Dept. of Dual Degree Master of Computer Applications, SNGIST College, Kerala, India

---***---

Abstract - *Data consolidation is a challenging issue in data integration. The usefulness of knowledge increases when it's linked and fused with other data from numerous (Web) sources. The promise of massive Data hinges upon addressing several big data integration challenges, like record linkage at scale, real-time data fusion, and integrating Deep Web. Although much work has been conducted on these problems, there's limited work on creating consistent, standard record from a gaggle of records like an equivalent real-world entity. We refer to this task as record normalization. Such a record representation, coined normalized record, is vital for both front-end and back-end applications. In this paper, we propose the record normalization complication, in attendance with in-depth analysis of normalization levels and of normalization structure. We offer an encyclopedic chassis for computing the normalized record. The suggested framework incorporate a outfit of record normalization procedures, from callow ones, which use only the mastery congregated from records themselves, to nexus master plan, which world widely dig a cluster of identical records before choosing an aid for an accredit 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 the ones to be used in practice.*

Key Words: **Normalization, Consolidation, Management Information System, Dataset, Duplicate Records, Algorithm.**

1. INTRODUCTION

Processing and analysing different types of data is Indispensable in the process of modelling and forecasting Of events' development, analysis of situations, defining Growth strategies, as well as decision support systems. The peculiarity of the current science research is the need to conduct of analysis not only the data various types, but also their semantics. The active development of means of operative collection and processing of the data various types, loading them into a knowledge base of decision making support system, analysing and forecasting is inherent in the energy sector, the medical sector, the financial market, the decision-making sector. The main problems that arise when processing of the data various types (the object being studied - is numeric data, image data, poorly structured reports, etc.), there is a rapid increase in the volume of data collected, the lack of methods for their effective analysis, the need for significant human resources to support of the data analysis Thus, the process of data consolidation, namely, the data of different origins of large size, in relation to analysis and prediction of the object's behaviour of the studied area Requires solving a number of problems:

->Increasing the efficiency of obtaining, analysing and using the information necessary to support decision making on determining the object's state;

->Improving the quality of decision-making forecasting through operation of information obtained from a reliable Source.

->Identification of new aspects of the object activity through analysis of data that was not foreseen and not taken into account when making decisions;

->Elimination of negative tendencies and undesirable consequences for forecasting changes in the state of being investigated object, with timely detection.

There is a need to construct a model of consolidated heterogeneous data of the object, which in aggregate are endowed with signs of completeness, integrity, consistency and constitute an adequate the object information model of the investigated area, with a view to its analysis of processing and effective use in decision making support processes.

2. RELATED WORK

In this section, we review the literature on record normalization. We give a few pointers on the related problems of schema integration and ontology merging.

The problem of normalization of database records Was first described by Culotta et al. [1]. They provided the first attempt to formalize the record normalization problem and proposed three solutions. The first solution uses string edit distance to determine the most central record. The second solution optimizes the edit distance parameters, and the third one describes a feature-based solution to improve performance by means of a knowledge base. Their approach is an instance of typical field value normalization. They did not consider value-component-level normalization. In addition, their gold standard dataset has many instances of unreasonable normalized records.

Swoosh [2] describes record Merge operator, however, the purpose of the operator is not for producing normalized records, but rather for improving the ability to establish difficult record matching.

Wicketal.[3]proposed a discriminatively trained model to implement schema matching, reference, and normalization jointly. But the complexity of the model is greatly increased. This paper also contains no discussion on complete normalization at the Value-component level. Besides the above works that explicitly address record normalization, a few others include (or refer to) the general idea of record normalization in some form.

Tejada et al. [4] devise a system too automatically Extract and consolidate information from multiple sources into a unified database. Although object de-duplication is the primary goal of their research, record normalization arises when the system presents results to the user. They propose Ranking the strings for each attribute based on the user's confidence in the data source from which the string was extracted.

Wang et al. [5] propose a hybrid framework for upshot normalization in web merchandise by schema integration and data cleaning. Although their work mainly focuses on record matching, they consider the matter of filling missing data and repairing incorrect data, which has relevancy to record normalization.

Chaturvedi et al. [6] propose an automatic pattern discovery method for rule-based data standardization systems. Their goal is to help domain expert find the important and prevalent patterns for rule writing. Although they do not directly explore the problem of record normalization, their pattern discovery approach could be used for complete normalization. Label normalization in schema integration is related to record normalization.

Dragut et al. [7] propose a naming framework to assign meaningful labels to the elements of an integrated query interface. Their approach can capture the consistency among the labels assigned to various attributes within a global interface.

Ontology merging is another area related to record normalization [8]. A domain expert is routinely deeply heavily necessitate throughout the merging process, as long as our perspective endeavor to lessen human participation as much as possible.

Nataliia Melnykova[9] propose that the peculiarity of the consolidation of heterogeneous data of the object under study consists of the following steps:

-> Analysis of the information that have to be found;

->Analysis of the semantic values of entities and attributes;

-> Specification of semantic correspondences - by using of consolidated data and data dictionaries, and also defines the missing links between concepts.

-> The construction of a single meta-model, based on correspondences defined at previous stages, and differences in the data structures, consists in the formation of the structure of the warehouse of consolidated data.

->The output of the resulting mappings between entities and attributes is actually "integration", the virtual transfer of data from sources to the consolidated data warehouse.

Semantic relationships between data sources are unknown beforehand. Dependence is established using a detailed description of the source structure (meta-model) and its comparison with the meta-model of the state's space. Also, a data dictionary is used to identify names - synonyms of the object characteristics.

3. PROPOSED SYSTEM

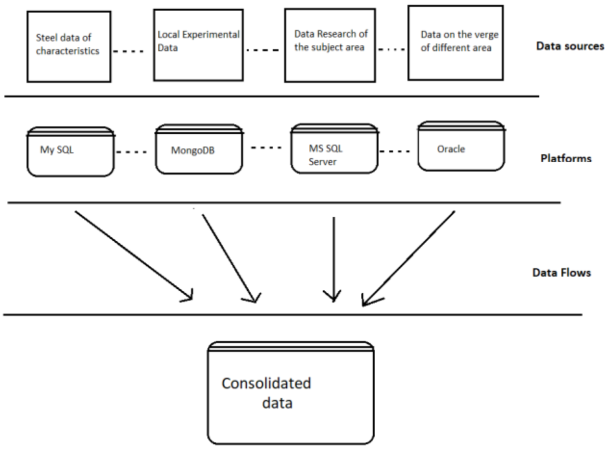

Fig 1: Consolidation of data

In the following sections, we give the details of our key techniques: (1) ranking-based strategies, (2) value component mining, and (3) ranked list merging.

3.1 Ranking-based Strategies

We utilize four ranking strategies: frequency, length, centroid, and feature-based. We use them to construct several rankers at record and field levels. To give a uniform presentation, we refer to records and their fields as units in this section. Let U be a bag of units for the same entity e. (It is a bag because the same value or the same record may appear multiple times.) U has p distinct units denoted by,

U = {u1,...,up}

If a ranker γ ranks a unit u higher than another unit v then we interpret this as saying that u is more appropriate as a normalized unit than v, according to γ.

3.1.1 Frequency Ranker

This ranker is defined as the ordered list of distinct units FR(U) = [u1,...,up], where ui appears more frequently than uj in U, for i < j.

3.1.2 Length Ranker

Length ranker is defined as the ordered list of distinct units LR(U) = [u1,...,up], where the character length of ui is larger than that of uj, 1 ≤ i < j ≤ p.

3.1.3 Centroid Ranker

Let S M be a similarity measure between units. We define the unit centroid score of u ∈ U to be

UCS(u) =1 |U|2 ∑ v∈UαuαvSM(u,v)

where αu , αv denote the occurrence frequencies of u and v in U, respectively. The centroid ranker gives the ordered list of distinct units CR(U) = [u1,...,up], (4) where UCS(ui) ≥ UCS(uj), 1 ≤ i < j ≤ p.
We use three similarity measures for SM: edit-distance, bigram and Winkler Similarity.

3.1.4 Feature Based Ranker

Feature-based ranker is defined as the ordered list of distinct units
FBR(U) = [u1,...,up], (11)
where pn(ui) ≥ pn(uj), 1 ≤ i < j ≤ p. Let TS = {< v1,l1 > ... < v|TS|,l|TS| >}be a training set, where vi is the ith record in the training set. and li =1 if vi is the normalized unit of an entity, 0 otherwise.

The features for the feature-based rankers are as follows:

Strategy feature: These features are all binary, indicating if a unit is the first, second, or third highest ranked unit according to some strategy ranker.

Text features. We compute two features that examine the properties of the strings themselves. One is the acronym feature which is true if the matching unit contains a token in a list of known acronyms (e.g., "VLDB" in our running example)

3.2 Value Component Mining

We begin this section with a number of definitions to make the following description clear and consistent. Let V al(fj) = {ri[fj]|ri ∈ Re} be the collection of all values of the field fj among the records in Re.

The inverse document frequency(idf) of a term or a consecutive sequence of terms c is defined as idf(c,Re) = |Re| |{ri|ri ∈ Re,c ∈ ri[fj]}| (13) where|·| denotes set cardinality (the number of records in our case). Note that when c's frequency increases, c's idf decreases.

3.2.1 Mining Abbreviation-Definition Pairs

We propose an algorithm for this as follows:

Input: V al(fj) = {ri[fj]|ri ∈ Re} : the collection of all values of the field fj
Output: AWP: a set of abbreviation-word pairs

```
cwords = ∅; AWP = ∅;
pwords = tokenize(V al(fj))
uwords = unique(pwords);
for each uword ∈ uwords do
    if len(uword) ≥ ηlen andidf(uword,Re) ≤ ηidf then
        insert uword into cwords;
    end if
end for
for each cword ∈ cwords do
    pa words = getWordsBySameContext(cword,uwords,ηpos);
    if pa words/= ∅then
        abbreviations = getAbbreviations( cword,pa words);
    end if
    if abbreviations/= ∅then
        for each abbreviation ∈ abbreviations do
            insert (abbreviation,cword) into AWP;
        end for
    end if
end for
return AWP
```

3.2.2 Mining Template Collocations and Sub collocations

Input: CV al(fj) – the updated version of V al(fj) with abbreviations extended by Algorithm 1.
Input: ηidf.
Output: TCSP: a set of pairs{(tc,Stc)}, where tc is a template collocation and Stc its subcollocations.

```
TCSP = ∅; m=getMaxWordCount(CV al(fj));
1-collocs = getOneWordCollocations(CV al(fj));
if 1-collocs == ∅then
    return ∅
end if
for each 1-colloc ∈ 1-collocs do
    add (1-colloc,∅) to TCSP;
end for
ews = getCandidateExpandWords(CV al(fj))); //Rule 1
for n = 2 to m do
    n-collocs = getNCollocations(CV al(fj),n,ηidf);
    if n-collocs == ∅then
        break;
    end if
    Y = ∅; //pairs to be ignored
    for each n-colloc∈n-collocs do
        cspairs = getExpandedSubcollocationPairs( n-colloc, ews, TCSP);
        if cspairs/= ∅then
            for each cspair ∈ cspairs do
                {cspair is of the form (c,Sc), c is a collocation and Sc its set of subcollocations; c is a subcollocation of n-colloc}
                X ={c}∪Sc;
                insert (n-colloc, X) into TCSP;
                add cspair to Y ; //not a template collocation
            end for
        end if
    end for
    TCSP = TCSP −Y ;
end for
remove the pairs of the form (c,∅) from TCSP;
return TCSP
```

3.2.3 Frequent Template Collocation Mining

The algorithm is as follows:

```
Input: CV al(fj) = {ri[fj]|ri ∈ Re}: the collection of all values of field fj
Input: ηtccr
Output: Tatwin: the set of most frequently co-occurring pairs of template collocations
Tatwin = ∅;
CV al(fj) = updateValWithAWP(V al(fj));
Z=MTS(V al(fj)); //Z has pairs of the form (tc,Stc)
TCj=getTemplateCollocations(Z); //TCj is the set of tc's
TCj=getTCPCounts(TCj,CV al(fj));
for each tc1 ∈ TCj do
    (tc2,ρ)=getMostFrequentTwinTC(tc1,TCj,CV al(fj));
    ρ2 = getCount(tc2, TCj);
    ratio= ρ ρ2 ;
    if ratio > ηtccr then
        insert (tc1,tc2) into Tatwin;
    end if
end for
return Tatwi
```

4. CONCLUSIONS

In this paper, we studied the problem of record normalization over a set of matching records that refer to the same real-world entity. We presented three levels of normalization granularities (record-level, field-level and value-component level) and two forms of normalization (typical normalization and complete normalization). For each form of normalization, we proposed a computational framework 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. For multi- strategy approach, we used result merging models inspired from meta-searching to mix the results from variety of single strategies. We analyzed the record and field level normalization in the typical normalization. In the complete normalization, we focused on field values and proposed algorithms for acronym expansion and value component mining to produce much improved normalized field values. We implemented a prototype and tested it on a real-world dataset. The experimental results demonstrate the feasibility and effectiveness of our approach. Our method outperforms the state-of-the-art by a significant margin. 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 currently has made this difficult. Second, investigate how to add an effective human-in-the-loop component into the current solution as automated solutions alone won't be able to achieve perfect accuracy.

REFERENCES

[1] A. Culotta, M. Wick, R. Hall, M. Marzilli, and A. McCallum, "Canonicalization of database records using adaptive similarity measures," in SIGKDD, 2007, pp. 201–209.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[2] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom, "Swoosh: A generic approach to entity resolution," VLDBJ, vol. 18, no. 1, pp. 255–276, 2009

[3] M. L. Wick, K. Rohanimanesh, K. Schultz, and A. McCallum,"unified approach for schema matching, coreference and canonicalization," in SIGKDD, 2008, pp. 722–730.

[4] S. Tejada, C. A. Knoblock, and S. Minton, "Learning object identification rules for information integration," Inf. Sys., vol. 26, no. 8, pp. 607–633, 2001.

[5] L. Wang, R. Zhang, C. Sha, X. He, and A. Zhou, "A hybrid framework for product normalization in online shopping," in DASFAA, vol. 7826, 2013, pp. 370–384.

[6] S. Chaturvedi and et al., "Automating pattern discovery for rule based data standardization systems," in ICDE, 2013.

[7] E. C. Dragut, C. Yu, and W. Meng, "Meaningful labeling of integrated query interfaces," in VLDB, 2006, pp. 679–690.

[8] S. Raunich and E. Rahm, "Atom: Automatic target-driven ontology merging," in ICDE, 2011, pp. 1276–1279.

[9] Nataliia Melnykova, Uliana Marikutsa, Uriy Kryvenchuk " The New Approaches of Heterogeneous Data Consolidation" IEEE September 2018

[10] Yongquan Dong, Eduard C. Dragut "Normalization of Duplicate Records from Multiple Resources" IEEE December 2018.