OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

NORMALIZATION OF IDENTICAL RECORDS FROM ONLINE RESOURCES

SHAIKH FARAZ SHAUKAT ALI

PG Student, Computer Science & Engineering Department, EESGOI, Aurangabad

---

Abstract: 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).We propose three levels of granularities for record normalization along with methods to construct normalized records according to them. We propose a comprehensive framework for systematic construction of normalized records. Our framework is flexible and allows new strategies to be added with ease. To our knowledge, this is the first piece of work to propose such a detailed framework. We propose and compare a range of normalization strategies, from frequency, length, centroid and feature-based to more complex ones that utilize result merging models from information retrieval, such as (weighted) Borda. We introduce a number of heuristic rules to mine desirable value components from a field. We use them to construct the normalized value for the field. We perform empirical studies on publication records. The experimental results show that the proposed weighted-Borda-based approach significantly outperforms the baseline approaches.

---∴∴∴---

I INTRODUCTION

The Web has evolved into a data-rich repository containing 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. We propose three levels of granularities for record normalization along with methods to construct normalized records according to them. 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 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. Record normalization is a challenging problem because different Web sources may represent the attribute values of an entity in different ways or even provide conflicting data. Conflicting data may occur because of incomplete data, different data representations, missing attribute values, and even erroneous data. They are extracted from different websites. Record Rnorm is constructed by hand for illustration purposes. One notices that the same publication has different representations in different websites. For instance, the eld author uses the format last-name, rst-name-initial in the record Ra, but the values of the same eld in the records Rb, Rc, and Rd use the format rst-name-initial. last-name. One can also observe that the value of the eld pages is absent in Ra. The field venue has incomplete values in three of the four records and has novaluein Rd; it contains the abbreviations proc, int, conf to represent proceedings, international and conference, respectively, in the records Ra and Rc; it contains the acronym VLDB to represent Very Large Data Bases while missing proceedings of the 32nd international conference on in Rb. Some values of the attributes of Rnorm cannot be acquired directly from the given set of matching records, such as the first names of the authors. They could be obtained by mining external sources, such as a search engine. We propose three levels of granularities for record normalization along

with methods to construct normalized records according to them. We propose a comprehensive framework for systematic construction of normalized records. Our framework is exible and allows new strategies to be added with ease. To our knowledge, this is the first piece of work to propose such a detailed framework. discovery have been performed and that the groups of true matching records have thus been identified. Our goal is to generate a uniform, standard record for each group of true matching records for end-user consumption. We call the generated record the normalized record. We call the problem of computing the normalized record for a group of matching records the record normalization problem (RNP), and it is the focus of this work. RNP is another specific interesting problem in data fusion. Record normalization is important in many application domains.

II LITERATURE SURVEY

1. K. C.-C. Chang and J. Cho, Accessing the web: From search to integration, in SIGMOD, 2006, pp. 804805.

In this paper, we formulated and solved the query planning and optimization problem for deep web databases with dependencies. We have developed a dynamic query planner with an approximation algorithm with a provable approximation ratio of 1/2. We have also developed cost models to guide the planner. The query planner automatically selects best sub-goals on-the-fly. The K query plans generated by the planner can provide alternative plans when the optimal one is not feasible. Our experiments show that the cost model for query planning is effective. Despite using an approximate algorithm, our planning algorithm outperforms the naive planning algorithm, and obtains the optimal query plans for most experimental queries in terms of both number of databases involved and actual execution time. We also show that our system has good scalability

2. Michael J. Cafarella, "WebTables: Exploring the Power of Tables on the Web"

In this paper We described the WebTables system, which is the first largescale attempt to extract and leverage the relational information embedded in HTML tables on the Web. We described how to support effective search on a massive collection of tables and demonstrated that current search engines do not support such search effectively. Finally, we showed that the recovered relations can be used to create what we believe is a very valuable data resource,
the attribute correlation statistics database

3. Ju Fan, "A Hybrid Machine-Crowdsourcing System for Matching Web Tables"

In this paper We have described a hybrid machine-crowdsourcing framework to effectively discover the schema matches for web tables. To the best of our
knowledge, our system is the first hybrid machine-crowdsourcing system for tackling the web table matching problem. Unlike traditional schema matching techniques, the machine part of our framework leverages a concept based approach. Due to the inherent semantic heterogeneity in web tables, pure machine algorithms cannot always work well. To this end, we harness the power of human intelligence as the crowd sourcing part of our framework to further improve the matching quality

4. Xin Luna Dong,"Big Data Integration"

This seminar reviews the state-of-the-art techniques for data integration in addressing the new challenges raised by Big Data, including volume and number of sources, velocity, variety, and veracity. We discuss how close we are to meeting these challenges and identify many open problems for future research.

5. Kyosuke Nishida, Kugatsu Sadamitsu, "Understanding the Semantic Structures of Tables with a Hybrid Deep Neural Network Architecture"

In this paper We proposed a new deep neural network architecture, TabNet, for table type classification, where six table types are defined on the basis of semantic triples of the form (subject, property, object) that the tables contain. Our architecture reflects the structure of tables: each cell has a sequence of tokens, and a table is a matrix of cells. It consists of an RNN that encodes token sequences and a CNN that extracts semantic features, e.g., the existence of rows describing properties, to classify table types

III SYSTEM ARCHITECTURE

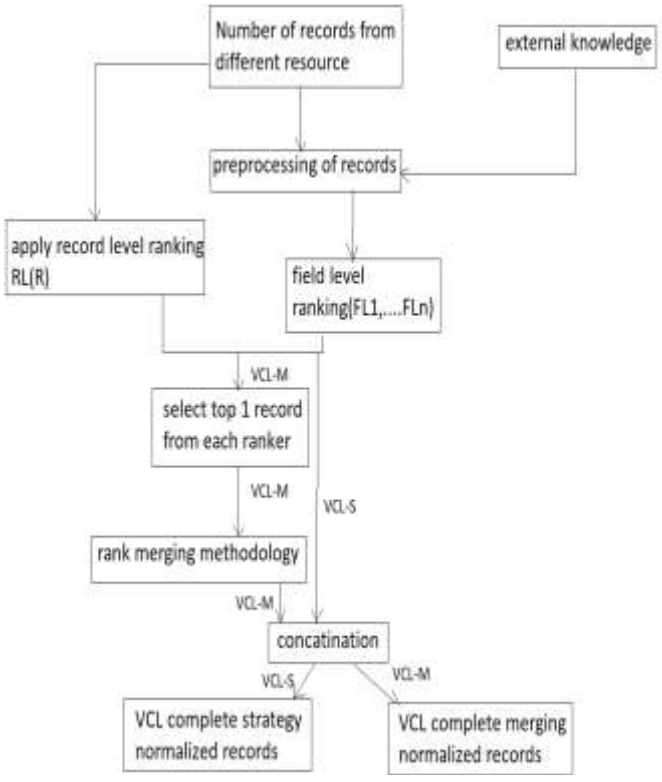

Figure 1 : System Architecture

In this system, 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 introduce a number of heuristic rules to mine desirable value components from a field. We use them to construct the normalized value for the field. We perform empirical studies on publication records. The experimental results show that the proposed weighted-Bordabased approach significantly outperforms the baseline approaches. We propose three levels of granularities for record normalization along with methods to construct normalized records according to them. We propose a comprehensive framework for systematic construction of normalized records. Our framework is flexible and allows new strategies to be added with ease. To our knowledge, this is the first piece of work to propose such a detailed framework. We propose and compare a range of normalization strategies, from frequency, length, centroid and feature-based to more complex ones that utilize result merging models from information retrieval, such as (weighted) Borda.

IV RESULTS

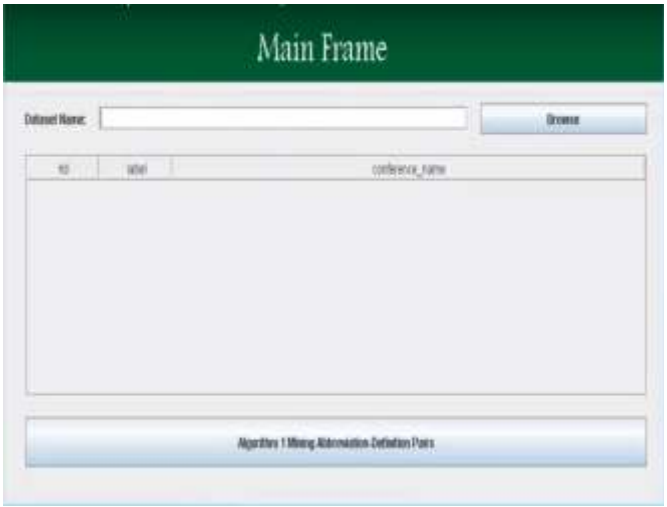

Figure 2: Home Page

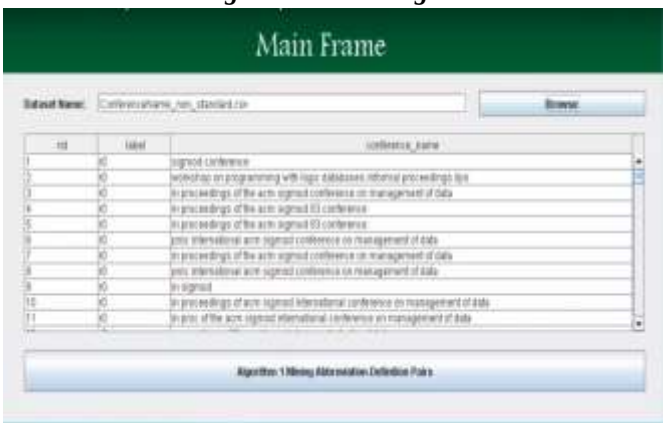

Figure 3: Load Dataset

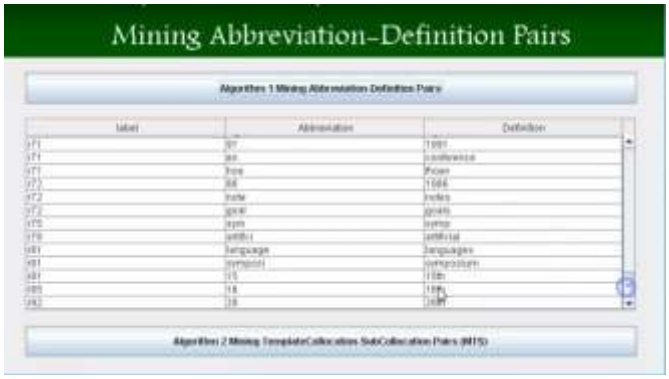

Figure 4: Algorithm 1 Mining Abbreviation-Definition Pairs

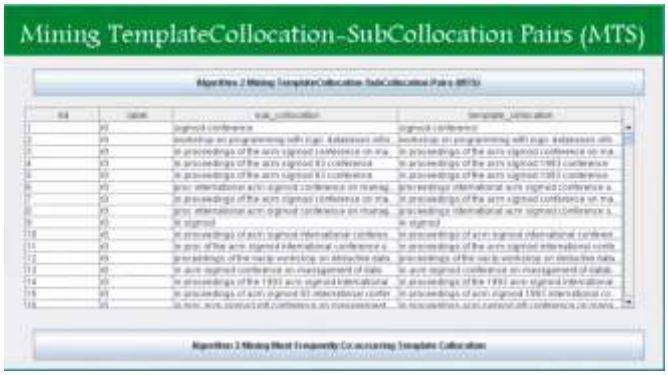

Figure 5: Algorithm 2 Mining Template Collocation-SubCollection Pairs(MTS)

V CONCLUSION

In this system, 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 introduce a number of heuristic rules to mine desirable value components from a field. We use them to construct the normalized value for the field. We perform empirical studies on publication records. We propose and compare a range of normalization strategies, from frequency, length, centroid and feature-based to more complex ones that utilize result merging models from information retrieval, such as (weighted) Borda.

REFERENCES

1. K. C.-C. Chang and J. Cho, Accessing the web: From search to integration, in SIGMOD, 2006, pp. 804805.
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. 538549, 2008.
3. W. Meng and C. Yu, Advanced Metasearch Engine Technology. Morgan & Claypool Publishers, 2010.
4. A. Gruenheid, X. L. Dong, and D. Srivastava, Incremental record linkage, PVLDB, vol. 7, no. 9, pp. 697708, 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. 4253.

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. Kopcke and E. Rahm, Frameworks for entity matching: A comparison, DKE, vol. 69, no. 2, pp. 197210, 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. 116, 2007
10. W. Su, J. Wang, and F. Lochovsky, Record matching over query results from multiple web databases, TKDE, vol. 22, no. 4, 2010.
11. H. Kopcke and E. Rahm, Frameworks for entity matching: A comparison, DKE, vol. 69, no. 2, pp. 197210, 2010.
12. X. Yin, J. Han, and S. Y. Philip, Truth discovery with multiple conicting information providers on the web, ICDE, 2008.
13. A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, Duplicate record detection: A survey, TKDE, vol. 19, no. 1, pp. 116, 2007.
14. P. Christen, A survey of indexing techniques for scalable record linkage and deduplication, TKDE, vol. 24, no. 9, 2012.
15. S. Tejada, C. A. Knoblock, and S. Minton, Learning object identication rules for information integration, Inf. Sys., vol. 26, no. 8, pp. 607633, 2001.
16. L. Shu, A. Chen, M. Xiong, and W. Meng, Efcient spectral neighborhood blocking for entity resolution, in ICDE, 2011.
17. Y. Jiang, C. Lin, W. Meng, C. Yu, A. M. Cohen, and N. R. Smalheiser, Rulebased deduplication of article records from bibliographic databases, Database, vol. 2014, 2014.
18. X. Li, X. L. Dong, K. Lyons, W. Meng, and D. Srivastava, Truth nding on the deep web:Is the problem solved inPVLDB,vol.6, no. 2, 2012, pp. 97108.
19.J.PasternackandD.Roth,Makingbetterinformedtrustdecisions with generalized fact-nding, in IJCAI, 2011, pp. 23242329.
20. X. L. Dong and F. Naumann, Data fusion: resolving data conicts for integration, PVLDB, vol. 2, no. 2, pp. 16541655, 2009.
21. E. K. Rezig, E. C. Dragut, M. Ouzzani, A. K. Elmagarmid, and W.G.Aref
22. X. Wang, X. L. Dong, and A. Meliou, Data x-ray: A diagnostic tool for data errors, in SIGMOD, 2015, pp. 12311245.
23. G. R. D. Patrick AV Hall, Approximate string matching, ACM Computing Surveys, vol. 12, no. 4, pp. 381402, 1980.
24. 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. 7378.
25. D. C. Liu and J. Nocedal, On the limited memory bfgs method for large scale optimization, Mathematical Programming, vol. 45, no. 3, pp. 503528, 1989.