Trends in Cleaning Relational Data: Consistency and Deduplication

Ihab F. Ilyas
University of Waterloo
ilyas@uwaterloo.ca

Xu Chu
University of Waterloo
x4chu@uwaterloo.ca

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Databases

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

I. F. Ilyas and X. Chu. *Trends in Cleaning Relational Data: Consistency and Deduplication.* Foundations and Trends® in Databases, vol. 5, no. 4, pp. 281–393, 2012.

This Foundations and Trends® issue was typeset in LaTeX using a class file designed by Neal Parikh. Printed on acid-free paper.

ISBN: 978-1-68083-023-1

Full text available at: http://dx.doi.org/10.1561/1900000045

Foundations and Trends® in Databases

Volume 5, Issue 4, 2012

Editorial Scope

Topics

Foundations and Trends® in Databases covers a breadth of topics relating to the management of large volumes of data. The journal targets the full scope of issues in data management, from theoretical foundations, to languages and modeling, to algorithms, system architecture, and applications. The list of topics below illustrates some of the intended coverage, though it is by no means exhaustive:

- Data models and query languages
- Query processing and optimization
- Storage, access methods, and indexing
- Transaction management, concurrency control, and recovery
- Deductive databases
- Parallel and distributed database systems
- Database design and tuning
- Metadata management
- Object management
- Trigger processing and active databases
- Data mining and OLAP
- Approximate and interactive query processing
- Data warehousing
- Adaptive query processing
- Data stream management
- Search and query integration
- XML and semi-structured data
- Web services and middleware
- Data integration and exchange
- Private and secure data management
- Peer-to-peer, sensornet, and mobile data management
- Scientific and spatial data management
- Data brokering and publish/subscribe
- Data cleaning and information extraction
- Probabilistic data management

Information for Librarians

Foundations and Trends® in Databases, 2012, Volume 5, 4 issues. ISSN paper version 1931-7883. ISSN online version 1931-7891. Also available as a combined paper and online subscription.

Foundations and Trends® in Databases
Vol. 5, No. 4 (2012) 281–393
© 2015 I. F. Ilyas and X. Chu
DOI: 10.1561/1900000045

Trends in Cleaning Relational Data: Consistency and Deduplication

Ihab F. Ilyas
University of Waterloo
ilyas@uwaterloo.ca

Xu Chu
University of Waterloo
x4chu@uwaterloo.ca

Full text available at: http://dx.doi.org/10.1561/1900000045

Contents

1 Introduction **2**

1.1 Notations 4

1.2 Outline 6

2 Taxonomy of Anomaly Detection Techniques **7**

2.1 What to Detect 9

2.2 How to Detect 37

2.3 Where to Detect 42

3 Taxonomy of Data Repairing Techniques **50**

3.1 What to Repair 52

3.2 How to Repair 71

3.3 Where to Repair 86

4 Big Data Cleaning **94**

5 Conclusion **100**

References **103**

Abstract

Data quality is one of the most important problems in data management, since dirty data often leads to inaccurate data analytics results and wrong business decisions. Poor data across businesses and the government cost the U.S. economy $3.1 trillion a year, according to a report by InsightSquared in 2012.

To detect data errors, data quality rules or integrity constraints (ICs) have been proposed as a declarative way to describe legal or correct data instances. Any subset of data that does not conform to the defined rules is considered erroneous, which is also referred to as a violation.

Various kinds of data repairing techniques with different objectives have been introduced, where algorithms are used to detect subsets of the data that violate the declared integrity constraints, and even to suggest updates to the database such that the new database instance conforms with these constraints. While some of these algorithms aim to minimally change the database, others involve human experts or knowledge bases to verify the repairs suggested by the automatic repeating algorithms.

In this paper, we discuss the main facets and directions in designing error detection and repairing techniques. We propose a taxonomy of current anomaly detection techniques, including error types, the automation of the detection process, and error propagation. We also propose a taxonomy of current data repairing techniques, including the repair target, the automation of the repair process, and the update model. We conclude by highlighting current trends in "big data" cleaning.

I. F. Ilyas and X. Chu. *Trends in Cleaning Relational Data: Consistency and Deduplication*. Foundations and Trends® in Databases, vol. 5, no. 4, pp. 281–393, 2012.
DOI: 10.1561/1900000045.

1

Introduction

As businesses generate and consume data more than ever, enforcing and maintaining the quality of their data assets become critical tasks. One in three business leaders does not trust the information used to make decisions [36], since establishing trust in data becomes a challenge as the variety and the number of sources grow. For example, in health care domains, inaccurate or incorrect data may threaten patient safety [75].

Gartner predicted that more than 25% of critical data in the world's top companies is flawed [106]. Poor data across businesses and the government costs the U.S. economy $3.1 trillion a year, according to a report by InsightSquared [29]. With the increasing popularity of data science, it became evident that data curation, preparation, cleaning, and other "janitorial" data tasks, are key enablers in unleashing value of data, as indicated in a 2014 article in the New York Times[1].

Even when the data is ingested in JSON, XML, or text format, many of data quality assessment and cleaning activities happen after transforming the data into relational tables. There are many notions related to relational data quality: data consistency, data accuracy, data completeness, and data currency. Data consistency refers to the valid-

[1] `http://nyti.ms/1t8IzfE`

ity and integrity of data; data accuracy refers to how accurate the data values in a database with respect to the true values; data completeness indicates whether all the data needed to meet the information needs is available; and data currency, also known as, data timeliness, gives the degree to which the data is current with respect to the world or the process it models. There are various surveys and books on relational data quality. Rahm and Do [93] give a classification of different types of errors that can happen in an Extract-Transform-Load (ETL) process, and survey the tools available for cleaning data in an ETL process; some focus on the effect of incompleteness data on query answering [61], and the use of a Chase procedure for dealing with incomplete data [62]; Hellerstein [67] focuses on cleaning quantitative data, such as integers and floating points, using mainly statistical outlier detection techniques. Bertossi [8] provides complexity results for repairing inconsistent data, and performing consistent query answering on inconsistent data; Fan and Geerts [44] discuss the use of data quality rules in data consistency, data currency, and data completeness, how different aspects of data quality issues might interact; and Dasu and Johnson [33] summarize how techniques in exploratory data mining can be integrated with data quality management.

In this paper, we focus on the data consistency aspect of relational data quality. To ensure data consistency, data quality rules are often used. We use integrity constraints (ICs) to express data quality rules. Any part of the data that does not conform to a given set of ICs is considered erroneous, also known as a violation of ICs. Data deduplication can be seen as enforcing a key constraint defined on all the attributes of a relational schema, since two duplicate tuples can be seen as a violation of the key constraint. Data cleaning, in this context, is the exercise of detecting errors, and possibly modifying the database, such that the data conforms to a set of data quality rules expressed in a variety of languages. This paper covers techniques to detect data inconsistencies, as well as techniques to repair data inconsistencies.

The following example illustrates a real world tax record database that has various data quality problems due to the violations of different data quality rules, and the existence of duplicate records.

Example 1.1. Consider the US tax records in Table 1.1. Each record describes an individual's address and tax information with 15 attributes: first and last name (FN, LN), gender (GD), area code (AC), mobile phone number (PH), city (CT), state (ST), zip code (ZIP), marital status (MS), has children (CH), salary (SAL), tax rate (TR), tax exemption amount if single (STX), married (MTX), and having children (CTX).

The following constraints hold: (1) area code and phone identify a person; (2) two persons with the same zip code live in the same state; (3) a person who lives in Los Angeles lives in California; (4) if two persons live in the same state, the one with lower salary has a lower tax rate; (5) tax exemption is less than the salary.

A violation with respect to an IC is defined as the minimal subset of database cells, such that at least one of the cells has to be modified to satisfy the IC, where a cell is an attribute value of a tuple, *e.g.*, Cell $t_1[\text{FN}]$ corresponds to Attribute FN of Tuple t_1 . For instance, the set of four cells $\{t_1[\text{ZIP}], t_8[\text{ZIP}], t_1[\text{ST}], t_8[\text{ST}]\}$ is a violation with respect to the second constraint. Furthermore, Record t_4 and t_9 refer to the same person, even though $t_4[\text{FN}]$ and $t_9[\text{FN}]$ are different, and $t_9[\text{AC}]$ is empty. Given a relational database instance I of schema R and a set of integrity constraints Σ, we need to find another database instance I' with no violations with respect to Σ.

1.1 Notations

Let R denote a relational schema, and I be an instance of that schema. Attributes of R are denoted as $attr(R) = \{A_1, \ldots, A_m\}$. For each Attribute A in R, let $Dom(A)$ denote the domain of A. I consists of a set of tuples, each of which belongs to the domain $Dom(A_1) \times \ldots \times Dom(A_m)$. We assume that there is a unique tuple identifier associated with each tuple $t \in I$. Let $TIDs(I)$ denote the set of all tuple identifiers. We identify a cell of Attribute A of a tuple t in I by $I(t[A])$, simply referred to as $t[A]$ when the context is clear. Let $CIDs(I)$ denote the set of all cell identifiers in I.

TID	FN	LN	GD	AC	PH	CT	ST	ZIP	MS	CH	SAL	TR	STX	MTX	CTX
t_1	Mark	Ballin	M	304	232-7667	Anthony	WA	25813	S	Y	70000	3	2000	0	2000
t_2	Chunho	Black	M	206	154-4816	Seattle	WA	98103	M	N	60000	4.63	0	0	0
t_3	Annja	Rebizant	F	636	604-2692	Cyrene	MO	64739	M	N	40000	6	0	4200	0
t_4	Annie	Puerta	F	501	378-7304	West Crossett	AR	72045	M	N	85000	7.22	0	40	0
t_5	Anthony	Landram	M	319	150-3642	Gifford	IA	52404	S	Y	15000	2.48	40	0	40
t_6	Mark	Murro	M	970	190-3324	Denver	CO	80251	S	Y	60000	4.63	0	0	0
t_7	Ruby	Billinghurst	F	501	154-4816	Kremlin	AR	72045	M	Y	70000	7	0	35	1000
t_8	Marcelino	Nuth	F	304	540-4707	Kyle	WV	25813	M	N	10000	4	0	0	0
t_9	Ann	Puerta	F		378-7304	West Crossett	AR	72045	M	N	86000	7.22	0	40	0

Table 1.1: Tax data records.

1.2 Outline

The remainder of the paper is organized as follows. Section 2 discusses different ways to detect anomalies in the data, such as data duplication, integrity constraints languages, along with algorithms for their automatic discovery, and provenance-based error propagation, based on what, how, and where to detect. Section 3 introduces the taxonomy we adopt to classify data repairing techniques, based on what, how, and where to repair, and presents the details of multiple techniques in each dimension. Section 4 discusses the techniques proposed for dealing with big data cleaning. Section 5 concludes and summarizes future research directions.

References

[1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases.* Addison-Wesley, 1995.

[2] F. N. Afrati and P. G. Kolaitis. Repair checking in inconsistent databases: algorithms and complexity. In *Proceedings of the 12th International Conference on Database Theory*, pages 31–41, 2009.

[3] N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: ranking and clustering. *Journal of the ACM (JACM)*, 55(5):23, 2008.

[4] R. Ananthakrishna, S. Chaudhuri, and V. Ganti. Eliminating fuzzy duplicates in data warehouses. In *Proceedings of the 28th international conference on Very Large Data Bases*, pages 586–597, 2002.

[5] A. Arasu, M. Götz, and R. Kaushik. On active learning of record matching packages. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 783–794, 2010.

[6] A. Arasu, C. Ré, and D. Suciu. Large-scale deduplication with constraints using dedupalog. In *Proceedings of the 2009 IEEE International Conference on Data Engineering*, pages 952–963, 2009.

[7] M. Arenas, L. Bertossi, and J. Chomicki. Consistent query answers in inconsistent databases. In *Proceedings of the 18th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 68–79, 1999.

[8] L. E. Bertossi. *Database Repairing and Consistent Query Answering.* Morgan & Claypool Publishers, 2011.

[9] G. Beskales, I. F. Ilyas, and L. Golab. Sampling the repairs of functional dependency violations under hard constraints. *Proceedings of the VLDB Endowment*, 3(1-2):197–207, 2010.

[10] G. Beskales, I. F. Ilyas, L. Golab, and A. Galiullin. On the relative trust between inconsistent data and inaccurate constraints. In *29th IEEE International Conference on Data Engineering*, pages 541–552, 2013.

[11] G. Beskales, I. F. Ilyas, L. Golab, and A. Galiullin. Sampling from repairs of conditional functional dependency violations. *The VLDB Journal*, 23(1):103–128, 2014.

[12] G. Beskales, M. A. Soliman, I. F. Ilyas, and S. Ben-David. Modeling and querying possible repairs in duplicate detection. *Proceedings of the VLDB Endowment*, pages 598–609, 2009.

[13] I. Bhattacharya and L. Getoor. Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):5, 2007.

[14] M. Bilenko, B. Kamath, and R. J. Mooney. Adaptive blocking: Learning to scale up record linkage. In *6th International Conference on Data Mining*, pages 87–96, 2006.

[15] M. Bilenko and R. J. Mooney. Adaptive duplicate detection using learnable string similarity measures. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 39–48. ACM, 2003.

[16] J. Bleiholder and F. Naumann. Data fusion. *ACM Computing Survey*, 41(1), 2008.

[17] P. Bohannon, W. Fan, M. Flaster, and R. Rastogi. A cost-based model and effective heuristic for repairing constraints by value modification. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, pages 143–154. ACM, 2005.

[18] P. Bohannon, W. Fan, F. Geerts, X. Jia, and A. Kementsietsidis. Conditional functional dependencies for data cleaning. In *Proceedings of the 23rd International Conference on Data Engineering*, pages 746–755, 2007.

[19] A. Chalamalla, I. F. Ilyas, M. Ouzzani, and P. Papotti. Descriptive and prescriptive data cleaning. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pages 445–456, 2014.

[20] S. Chaudhuri, B. Chen, V. Ganti, and R. Kaushik. Example-driven design of efficient record matching queries. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, pages 327–338, 2007.

[21] S. Chaudhuri, V. Ganti, and R. Motwani. Robust identification of fuzzy duplicates. In *Proceedings of the 21st International Conference on Data Engineering*, pages 865–876, 2005.

[22] F. Chiang and R. J. Miller. Discovering data quality rules. *Proceedings of the VLDB Endowment*, 1(1):1166–1177, 2008.

[23] F. Chiang and R. J. Miller. A unified model for data and constraint repair. In *Proceedings of the 2011 IEEE 27th International Conference on Data Engineering*, pages 446–457, 2011.

[24] J. Chomicki and J. Marcinkowski. Minimal-change integrity maintenance using tuple deletions. *Information and Computation*, 197(1):90–121, 2005.

[25] X. Chu, I. F. Ilyas, and P. Papotti. Discovering denial constraints. *Proceedings of the VLDB Endowment*, 6(13):1498–1509, 2013.

[26] X. Chu, I. F. Ilyas, and P. Papotti. Holistic data cleaning: Putting violations into context. In *29th IEEE International Conference on Data Engineering*, pages 458–469, 2013.

[27] X. Chu, I. F. Ilyas, P. Papotti, and Y. Ye. Ruleminer: Data quality rules discovery. In *IEEE 30th International Conference on Data Engineering*, pages 1222–1225, 2014.

[28] X. Chu, J. Morcos, I. F. Ilyas, M. Ouzzani, P. Papotti, N. Tang, and Y. Ye. KATARA: A data cleaning system powered by knowledge bases and crowdsourcing. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1247–1261, 2015.

[29] S. Clemens. 7 facts about data quality. *InsightSquared*, 2012.

[30] W. W. Cohen. Integration of heterogeneous databases without common domains using queries based on textual similarity. In *ACM SIGMOD Record*, volume 27, pages 201–212, 1998.

[31] G. Cong, W. Fan, F. Geerts, X. Jia, and S. Ma. Improving data quality: Consistency and accuracy. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, pages 315–326. VLDB Endowment, 2007.

[32] M. Dallachiesa, A. Ebaid, A. Eldawy, A. Elmagarmid, I. F. Ilyas, M. Ouzzani, and N. Tang. Nadeef: a commodity data cleaning system. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 541–552, 2013.

[33] T. Dasu and T. Johnson. *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons, Inc., 2003.

[34] F. De Marchi, S. Lopes, and J.-M. Petit. Unary and n-ary inclusion dependency discovery in relational databases. *Journal of Intelligent Information Systems*, 32(1):53–73, 2009.

[35] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[36] D. Deroos, C. Eaton, G. Lapis, P. Zikopoulos, and T. Deutsch. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill, 2011.

[37] T. Diallo, J.-M. Petit, and S. Servigne. Discovering editing rules for data cleaning. In *Proceedings of AQB conference*, page 40, 2012.

[38] A. Doan, Y. Lu, Y. Lee, and J. Han. Profile-based object matching for information integration. *IEEE Intelligent Systems*, 18(5):54–59, 2003.

[39] X. L. Dong and F. Naumann. Data fusion: resolving data conflicts for integration. *Proceedings of the VLDB Endowment*, 2(2):1654–1655, 2009.

[40] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16, 2007.

[41] M. Elsner and E. Charniak. You talking to me? a corpus and algorithm for conversation disentanglement. In *Association for Computational Linguistics (ACL)*, pages 834–842, 2008.

[42] M. Elsner and W. Schudy. Bounding and comparing methods for correlation clustering beyond ilp. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, pages 19–27, 2009.

[43] G. Fan, W. Fan, and F. Geerts. Detecting errors in numeric attributes. In *Web-Age Information Management*, pages 125–137. Springer, 2014.

[44] W. Fan and F. Geerts. *Foundations of Data Quality Management*. Synthesis Lectures on Data Management. 2012.

[45] W. Fan, F. Geerts, J. Li, and M. Xiong. Discovering conditional functional dependencies. *IEEE Transactions on Knowledge and Data Engineering*, 23(5):683–698, 2011.

[46] W. Fan, F. Geerts, S. Ma, and H. Müller. Detecting inconsistencies in distributed data. In *Proceedings of the 26th International Conference on Data Engineering*, pages 64–75, 2010.

[47] W. Fan, F. Geerts, N. Tang, and W. Yu. Inferring data currency and consistency for conflict resolution. In *29th IEEE International Conference on Data Engineering*, pages 470–481, 2013.

[48] W. Fan, F. Geerts, N. Tang, and W. Yu. Conflict resolution with data currency and consistency. *Journal of Data and Information Quality*, 5(1-2):6:1–6:37, 2014.

[49] W. Fan, X. Jia, J. Li, and S. Ma. Reasoning about record matching rules. *Proceedings of the VLDB Endowment*, 2(1):407–418, 2009.

[50] W. Fan, J. Li, S. Ma, N. Tang, and W. Yu. Towards certain fixes with editing rules and master data. *Proceedings of the VLDB Endowment*, 3(1-2):173–184, 2010.

[51] W. Fan, J. Li, S. Ma, N. Tang, and W. Yu. Interaction between record matching and data repairing. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, pages 469–480. ACM, 2011.

[52] W. Fan, J. Li, N. Tang, and W. Yu. Incremental detection of inconsistencies in distributed data. *IEEE Transactions on Knowledge and Data Engineering*, 26(6):1367–1383, 2014.

[53] H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C. Saita. Declarative data cleaning: Language, model, and algorithms. In *VLDB 2001, Proceedings of 27th International Conference on Very Large Data Bases*, pages 371–380, 2001.

[54] H. Galhardas, A. Lopes, and E. Santos. Support for user involvement in data cleaning. In *Data Warehousing and Knowledge Discovery - 13th International Conference, DaWaK 2011*, pages 136–151, 2011.

[55] F. Geerts, G. Mecca, P. Papotti, and D. Santoro. The llunatic data-cleaning framework. *Proceedings of the VLDB Endowment*, 6(9):625–636, 2013.

[56] F. Geerts, G. Mecca, P. Papotti, and D. Santoro. Mapping and cleaning. In *IEEE 30th International Conference on Data Engineering*, pages 232–243, 2014.

[57] F. Geerts, G. Mecca, P. Papotti, and D. Santoro. That's all folks! LLUNATIC goes open source. *Proceedings of the VLDB Endowment*, 7(13):1565–1568, 2014.

[58] L. Getoor and A. Machanavajjhala. Entity resolution: theory, practice & open challenges. *Proceedings of the VLDB Endowment*, 5(12):2018–2019, 2012.

[59] C. Gokhale, S. Das, A. Doan, J. F. Naughton, N. Rampalli, J. Shavlik, and X. Zhu. Corleone: Hands-off crowdsourcing for entity matching. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pages 601–612, 2014.

[60] L. Golab, H. Karloff, F. Korn, D. Srivastava, and B. Yu. On generating near-optimal tableaux for conditional functional dependencies. *Proceedings of the VLDB Endowment*, 1(1):376–390, 2008.

[61] G. Grahne. *The Problem of Incomplete Information in Relational Databases*, volume 554 of *Lecture Notes in Computer Science*. 1991.

[62] S. Greco, C. Molinaro, and F. Spezzano. Incomplete data and data dependencies in relational databases. *Synthesis Lectures on Data Management*, 2012.

[63] A. Gruenheid, X. L. Dong, and D. Srivastava. Incremental record linkage. *Proceedings of the VLDB Endowment*, 7(9):697–708, 2014.

[64] P. J. Guo, S. Kandel, J. Hellerstein, and J. Heer. Proactive wrangling: Mixed-initiative end-user programming of data transformation scripts. In *ACM User Interface Software & Technology (UIST)*, 2011.

[65] L. M. Haas, M. A. Hernández, H. Ho, L. Popa, and M. Roth. Clio grows up: from research prototype to industrial tool. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, pages 805–810, 2005.

[66] J. Heer, J. Hellerstein, and S. Kandel. Predictive interaction for data transformation. In *Conference on Innovative Data Systems Research (CIDR)*, 2015.

[67] J. M. Hellerstein. Quantitative data cleaning for large databases. *United Nations Economic Commission for Europe (UNECE)*, 2008.

[68] M. A. Hernández and S. J. Stolfo. The merge/purge problem for large databases. *ACM SIGMOD Record*, 24(2):127–138, 1995.

[69] M. A. Hernández and S. J. Stolfo. Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery*, 2(1):9–37, 1998.

[70] T. N. Herzog, F. J. Scheuren, and W. E. Winkler. *Data Quality and Record Linkage Techniques*. Springer Science & Business Media, 2007.

[71] Y. Huhtala, J. Kärkkäinen, P. Porkka, and H. Toivonen. TANE: An efficient algorithm for discovering functional and approximate dependencies. *Computer Journal*, 42(2):100–111, 1999.

[72] M. Interlandi and N. Tang. Proof positive and negative in data cleaning. In *31st IEEE International Conference on Data Engineering*, 2015.

[73] M. A. Jaro. Unimatch: A record linkage system: User's manual. *U.S. Bureau of the Census*, 1976.

[74] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Wrangler: Interactive visual specification of data transformation scripts. In *ACM Human Factors in Computing Systems (CHI)*, 2011.

[75] K. Kerr, T. Norris, and R. Stockdale. Data quality information and decision making: a healthcare case study. In *Proceedings of the 18th Australasian Conference on Information Systems Doctoral Consortium*, pages 5–7, 2007.

[76] Z. Khayyat, I. F. Ilyas, A. Jindal, S. Madden, M. Ouzzani, P. Papotti, J.-A. Quiané-Ruiz, N. Tang, and S. Yin. Bigdansing: A system for big data cleansing. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1215–1230, 2015.

[77] S. Kolahi and L. V. S. Lakshmanan. On approximating optimum repairs for functional dependency violations. In *12th International Conference on Database Theory*, pages 53–62, 2009.

[78] L. Kolb, A. Thor, and E. Rahm. Dedoop: efficient deduplication with hadoop. *Proceedings of the VLDB Endowment*, 5(12):1878–1881, 2012.

[79] L. Kolb, A. Thor, and E. Rahm. Load balancing for mapreduce-based entity resolution. In *IEEE 28th International Conference on Data Engineering*, pages 618–629, 2012.

[80] N. Koudas, A. Saha, D. Srivastava, and S. Venkatasubramanian. Metric functional dependencies. In *Proceedings of the 25th International Conference on Data Engineering*, pages 1275–1278, 2009.

[81] N. Koudas, S. Sarawagi, and D. Srivastava. Record linkage: similarity measures and algorithms. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, pages 802–803, 2006.

[82] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet Physics Doklady*, volume 10, page 707, 1966.

[83] A. Lopatenko and L. E. Bertossi. Complexity of consistent query answering in databases under cardinality-based and incremental repair semantics. In *11th International Conference on Database Theory*, pages 179–193, 2007.

[84] S. Ma, W. Fan, and L. Bravo. Extending inclusion dependencies with conditions. *Theoretical Computer Science*, 515:64–95, 2014.

[85] A. McCallum, K. Nigam, and L. H. Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the 6th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 169–178, 2000.

[86] A. McCallum and B. Wellner. Conditional models of identity uncertainty with application to noun coreference. In *Advances in Neural Information Processing Systems 17 Neural Information Processing Systems*, pages 905–912, 2004.

[87] A. Meliou, W. Gatterbauer, S. Nath, and D. Suciu. Tracing data errors with view-conditioned causality. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 505–516, 2011.

[88] M. Michelson and C. A. Knoblock. Learning blocking schemes for record linkage. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 440, 2006.

[89] A. E. Monge and C. Elkan. The field matching problem: Algorithms and applications. In *Knowledge Discovery and Data Mining*, pages 267–270, 1996.

[90] F. Naumann and M. Herschel. *An Introduction to Duplicate Detection.* Synthesis Lectures on Data Management. 2010.

[91] V. Ng and C. Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104–111, 2002.

[92] T. Papenbrock, S. Kruse, J.-A. Quiané-Ruiz, and F. Naumann. Divide & conquer-based inclusion dependency discovery. *Proceedings of the VLDB Endowment*, 8(7):774–785, 2015.

[93] E. Rahm and H. H. Do. Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23:2000, 2000.

[94] V. Raman and J. M. Hellerstein. Potter's wheel: An interactive data cleaning system. In *VLDB 2001, Proceedings of 27th International Conference on Very Large Data Bases*, pages 381–390, 2001.

[95] R. Russell. Index., Apr. 2 1918. US Patent 1,261,167.

[96] S. Sarawagi and A. Bhamidipaty. Interactive deduplication using active learning. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 269–278, 2002.

[97] A. D. Sarma, Y. He, and S. Chaudhuri. Clusterjoin: A similarity joins framework using map-reduce. *Proceedings of the VLDB Endowment*, 7(12):1059–1070, 2014.

[98] A. D. Sarma, A. Jain, A. Machanavajjhala, and P. Bohannon. An automatic blocking mechanism for large-scale de-duplication tasks. In *21st ACM International Conference on Information and Knowledge Management*, pages 1055–1064, 2012.

[99] G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *International Conference on Machine Learning*, pages 839–846, 2000.

[100] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52:55–66, 2010.

[101] P. Singla and P. Domingos. Entity resolution with markov logic. In *2013 IEEE 13th International Conference on Data Mining*, pages 572–582, 2006.

[102] S. Song and L. Chen. Discovering matching dependencies. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 1421–1424, 2009.

[103] W. M. Soon, H. T. Ng, and D. C. Y. Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544, 2001.

[104] B. Steuart and P. I. Staff. *The Daitch-Mokotoff Soundex Reference Guide*. Heritage Quest, 1994.

[105] M. Stonebraker, D. Bruckner, I. F. Ilyas, G. Beskales, M. Cherniack, S. B. Zdonik, A. Pagan, and S. Xu. Data curation at scale: The data tamer system. In *CIDR 2013, 6th Biennial Conference on Innovative Data Systems Research*, 2013.

[106] N. Swartz. Gartner warns firms of “dirty data”. *Information Management Journal*, 41(3), 2007.

[107] R. Taft. *Name Search Techniques*. Special report (New York State Identification and Intelligence System). Bureau of Systems Development, New York State Identification and Intelligence System, 1970.

[108] S. Tejada, C. A. Knoblock, and S. Minton. Learning object identification rules for information integration. *Information Systems*, 26(8):607–633, 2001.

[109] S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2:45–66, 2002.

[110] R. Vernica, M. J. Carey, and C. Li. Efficient parallel set-similarity joins using mapreduce. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 495–506. ACM, 2010.

[111] V. S. Verykios, A. K. Elmagarmid, and E. N. Houstis. Automating the approximate record-matching process. *Information Sciences*, 126(1):83–98, 2000.

[112] N. Vesdapunt, K. Bellare, and N. Dalvi. Crowdsourcing algorithms for entity resolution. *Proceedings of the VLDB Endowment*, 7(12), 2014.

[113] M. Volkovs, F. Chiang, J. Szlichta, and R. J. Miller. Continuous data cleaning. In *IEEE 30th International Conference on Data Engineering*, pages 244–255, 2014.

[114] J. Wang, T. Kraska, M. J. Franklin, and J. Feng. Crowder: Crowdsourcing entity resolution. *Proceedings of the VLDB Endowment*, 5(11):1483–1494, 2012.

[115] J. Wang, S. Krishnan, M. J. Franklin, K. Goldberg, T. Kraska, and T. Milo. A sample-and-clean framework for fast and accurate query processing on dirty data. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pages 469–480, 2014.

[116] J. Wang, G. Li, T. Kraska, M. J. Franklin, and J. Feng. Leveraging transitive relations for crowdsourced joins. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pages 229–240, 2013.

[117] J. Wang and N. Tang. Towards dependable data repairing with fixing rules. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pages 457–468. ACM, 2014.

[118] M. S. Waterman, T. F. Smith, and W. A. Beyer. Some biological sequence metrics. *Advances in Mathematics*, 20(3):367–387, 1976.

[119] M. Weis, F. Naumann, U. Jehle, J. Lufter, and H. Schuster. Industry-scale duplicate detection. *Proceedings of the VLDB Endowment*, 1(2):1253–1264, 2008.

[120] S. E. Whang and H. Garcia-Molina. Entity resolution with evolving rules. *Proceedings of the VLDB Endowment*, 3(1-2):1326–1337, 2010.

[121] W. E. Winkler. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. *Proceedings of the Section on Survey Research*, 1990.

[122] W. E. Winkler. The state of record linkage and current research problems. In *Statistical Research Division, U.S. Census Bureau*, 1999.

[123] E. Wu and S. Madden. Scorpion: Explaining away outliers in aggregate queries. *Proceedings of the VLDB Endowment*, 6(8):553–564, 2013.

[124] E. Wu, S. Madden, and M. Stonebraker. A demonstration of dbwipes: clean as you query. *Proceedings of the VLDB Endowment*, 5(12):1894–1897, 2012.

[125] C. M. Wyss, C. Giannella, and E. L. Robertson. FastFDs: A heuristic-driven, depth-first algorithm for mining functional dependencies from relation instances. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 101–110, 2001.

[126] M. Yakout, A. K. Elmagarmid, J. Neville, M. Ouzzani, and I. F. Ilyas. Guided data repair. *Proceedings of the VLDB Endowment*, 4(5):279–289, 2011.

[127] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica. Spark: cluster computing with working sets. In *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing*, volume 10, page 10, 2010.