Towards Reliable Data Analyses for Smart Cities

Tiago Brasileiro Araújo
Federal University of Campina Grande
Campina Grande, Brazil
tiagobrasileiro@copin.ufcg.edu.br

Cinzia Cappiello
Politecnico di Milano
Milan, Italy
cinzia.cappiello@polimi.it

Nadia Puchalski Kozievitch
Federal University of Technology
Curitiba, Brazil
nadiap@utfpr.edu.br

Demetrio Gomes Mestre
Federal University of Campina Grande
Campina Grande, Brazil
demetriogm@gmail.com

Carlos Eduardo Santos Pires
Federal University of Campina Grande
Campina Grande, Brazil
cesp@dsc.ufcg.edu.br

Monica Vitali
Politecnico di Milano
Milan, Italy
monica.vitali@polimi.it

ABSTRACT

As cities are becoming green and smart, public information systems are being revamped to adopt digital technologies. There are several sources (official or not) that can provide information related to a city. The availability of multiple sources enables the design of advanced analyses for offering valuable services to both citizens and municipalities. However, such analyses would fail if the considered data were affected by errors and uncertainties: Data Quality is one of the main requirements for the successful exploitation of the available information. This paper highlights the importance of the Data Quality evaluation in the context of geographical data sources. Moreover, we describe how the Entity Matching task can provide additional information to refine the quality assessment and, consequently, obtain a better evaluation of the reliability data sources. Data gathered from the public transportation and urban areas of Curitiba, Brazil, are used to show the strengths and effectiveness of the presented approach.

CCS CONCEPTS

• **Information systems → Geographic information systems**; *Data model extensions*; *Information integration*; *Entity resolution*; *Geographic information systems*;

KEYWORDS

Data Analysis, Entity Matching, Data Quality, Smart Cities

ACM Reference format:
Tiago Brasileiro Araújo, Cinzia Cappiello, Nadia Puchalski Kozievitch, Demetrio Gomes Mestre, Carlos Eduardo Santos Pires, and Monica Vitali. 2017. Towards Reliable Data Analyses for Smart Cities. In *Proceedings of IDEAS '17, Bristol, United Kingdom, July 12-14 2017,* 5 pages.
http://dx.doi.org/10.1145/3105831.3105834

IDEAS '17, July 12-14 2017, Bristol, United Kingdom

ACM ISBN 978-1-4503-5220-8/17/07…$15.00
http://dx.doi.org/10.1145/3105831.3105834

1 INTRODUCTION

The quality of life in a city greatly depends on the well-being of its citizens. To promote it, the municipalities invest in information systems that assist the citizens (e.g., regarding urban mobility or the identification of points of interest) directly or indirectly. In smart cities, the efficiency of public transportation and infrastructure investments can be improved by taking advantage of the big amount of data collected by municipalities. They enable the analysis of the condition of public transportation and of the feasibility of improvement actions that might require new solutions or modifications to the current infrastructure. Regarding urban areas (e.g., squares, gardens, hospitals, and parks), the municipalities can rationally allocate investments to infrastructure improvements. Since data are prone to errors and quality issues, it is important to assess Data Quality (DQ) before using them in order to take valuable strategical decisions. In fact, an analysis based on noisy or incomplete information can generate wrong results [6].

In this paper, we present a data quality assessment approach in which the application of Entity Matching (EM) aims to provide additional quality measures contributing to the data consistency of smart city information systems. The EM, also known as entity resolution, deduplication, record linkage, or reference reconciliation, is the task that identifies the same real-world object across different entity profiles [11]. In a smart city scenario, in which plenty of data sources are available (e.g., collaborative and official data sources), it is necessary to assume that multiple data sources can contain the same information with different data quality levels. Unfortunately, some of the consolidated EM techniques cannot be adopted since they mainly deal with traditional structured sources and in this context many sources are geographical datasets that describe urban areas (e.g., squares). The EM approach described in this paper addresses such issue by dealing with geographical data sources that present data overlapping.

2 DATA QUALITY ISSUES IN THE SMART CITY SCENARIO

Smart cities benefit of the data collected from multiple sources, such as official documentation, third parties information, environmental sensors, and so on, in order to improve the efficiency of public services as public transportation and infrastructure works. In the context of this work, the data sources belong to different

systems and have a static nature, i.e., besides coming from different repositories, they contain well defined and structured information that does not change over time or is updated with a low frequency. Examples of static data sources in a smart city scenario are urban areas, bus stops, timetables and, bus line trajectories. In this work, we consider the following static data sources of Curitiba: (i) Squares: contains 682 squares described by their names and geometrical descriptions and positioning; (ii) Bus Stops: contains approximately 6,982 georeferenced points distributed throughout the city.

The large amount of transportation data available enables the design of smart services that can support citizens and really change the way they live. However, in order to obtain the maximum value, it is necessary to consider that not all the data might be relevant: replicated data, missing or outdated values can negatively affect decisions [4]. Focusing on Curitiba, for example, the comparison of different official sources (such as IPPUC and Urbs) already listed several differences, such as 20% of the total of bus stops. If we compare the bus stops to non-official sources (such as Open Street Map), the differences are even bigger. For this reason, the EM task can be applied to measure the quality of data sources and provide information about the number of entity linkages between two data sources.

3 DATA QUALITY ANALYSIS

Acquiring knowledge about the sources of information is a prerequisite for getting valuable results from analytics application. Sources are usually subject to quality issues due to errors and missing information and assessing Data Quality can be a good starting point for identifying not significant information. In fact, data quality is able to identify and eliminate "noises" that can affect data interpretation and/or decisions. Data quality is often defined as *fitness for use*, i.e., the ability of a data collection to meet users' requirements [17] and it is evaluated by means of different dimensions (e.g., accuracy, completeness, timeliness and consistency), which definition mainly depends on the context of use (i.e.,data types, data sources and applications) [3]. For this reason, in the smart city scenario characterized by high source variety there is the need of adaptive approaches able to select the assessment algorithms to use on the basis of the data types and sources. In Figure 1, we propose an assessment architecture able to deal with different sources and users/application requirements. This architecture clarifies how the EM task can provide additional information to the data quality assessment. The *Data Quality Service Interface* lets the users and/or applications access the Data Quality service in order to gather metadata that describe the quality level of the analyzed data sources. Through this interface, they are also able to select data sources, filter data on the basis of data quality requirements (e.g., select only data objects with completeness greater than 80%) and choice data quality dimensions to evaluate. Their settings and preferences are saved for the subsequent invocations in the *Settings repository*.

All the dimensions and profile metadata are evaluated by the *DQ profiling and assessment* that is composed of two main components: the *Profiling module* and the *Assessment module*. The former is in charge to evaluate summaries and statistics about sources while the latter computes the data quality dimensions. Note that the assessment module triggers the most suitable algorithm to perform on

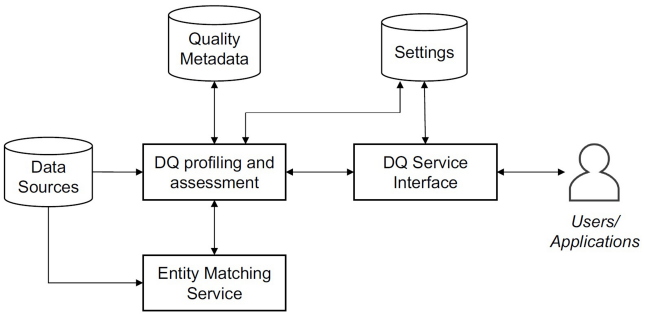

Figure 1: Data Quality assessment Architecture

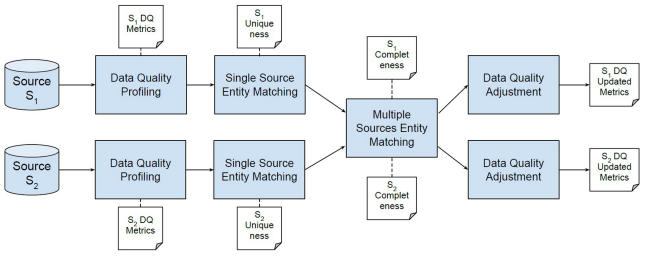

Figure 2: Data Quality enrichment through Entity Matching

the basis of the quality dimensions, on the type of values and on the users requirements. All the resulting values are stored in the *Quality Metadata* database. The Assessment module interacts with the *Entity Matching* service in order to gather additional information about the sources and refine the quality measures.

In fact, assessment procedures provide quality metadata of the different sources analyzing only their values. If data sources overlap, entity matching is needed: the same entity can be implicitly contained in several sources and the comparison of the different representation might reveal errors and inconsistencies and thus additional information about the quality of the sources. In this paper, we show that EM task can provide useful quality metrics (uniqueness and completeness) that aim to enrich the set of data quality metadata by providing more information about the reliability of a specific source, as described in Figure 2.

First of all, the EM supports the evaluation of the source accuracy by identifying duplicates. Redundancies need to be also identified among different sources and can help the evaluation of the completeness of each data source. Note that the Entity Matching task aims to solve the problem of identifying entities referring to the same real-world object [11]. The task is crucial in every information integration and data cleansing applications [10]. Given the pairwise-comparison nature of the task, EM is an intriguing problem that demands the proposition of new effective approaches in order to improve both the quality of similar entity pairs detection and its execution time.

In general, the problem of EM is defined by considering two data sources [6][13]. Given two sets of entities $A \in S$ and $B \in R$ from data sources S and R, the EM problem is to identify all correspondences between entities in the EM relation E. We denote

the schema of E as $R_E = (a_1, a_2, \ldots, a_n)$. Each a_i corresponds to an attribute (e.g., name, category, and geographic coordinate). An entity stored in the relation E assigns a value to each attribute. This means that the input data sources S and R contain a finite set of entities $e = [(a_1, v_1), (a_2, v_2), \ldots, (a_n, v_n)]$. Let $sim(e_1, e_2)$ be the similarity measure between entities e_1 and e_2, Φ_{max} the maximum threshold that defines whether e_1 matches e_2 and Φ_{min} the minimum threshold that defines the pair e_1 and e_2 as non-match. Thus, the task is to identify all pairs of entities $M = \{(e_i, e_k) \mid e_i, e_k \in S \text{ and } sim(e_i, e_k) \geq \Phi_{max}\}$, $NM = \{(e_i, e_k) \mid e_i, e_k \in S \text{ and } sim(e_i, e_k) \leq \Phi_{min}\}$ and $PM = \{(e_i, e_k) \mid e_i, e_k \in S \text{ and } \Phi_{min} < sim(e_i, e_k) < \Phi_{max}\}$ that are regarded as matches (M), non-matches (NM), and potential matches (PM), respectively. The definition includes the special case of finding similar entity pairs within a single source (i.e., $S = R$).

For a single source, the EM process (i) compares the entities in the source by using the Cartesian product, (ii) classifies the entity pairs and, from this information, (iii) determines the amount of duplicated entities contained in the data source. We use instead two data sources (R and S), in order to provide information about similar entities and, consequently, to support data integration. Thus, EM compares all entities of R with all entities of S.

3.1 Single data source

The EM task can be performed with all types of data. In particular, many EM algorithms are consolidated for structured data sets composed of text and numeric values (e.g., as proposed in [14]). In this work, we want to propose an Entity Matching (EM) approach[1] for dealing with entities that are represented by spatial geometries, e.g., polygons and points. Thus, the EM approach performs pairwise comparisons between geometries available at the data source and classifies each geometry pair according to the similarity between the involved geometries. To classify a geometry pair, linguistic and geographic matchers are applied. A linguistic matcher explores the textual attributes of the geometries (e.g., name and description) to determine the linguistic similarity (*LS*) of a geometry pair. To this end, algorithms such as Jaccard and Levenshtein distance are used. A geographic matcher explores the coordinates of both geometries in order to identify the overlapping area or the distance between them. The proportion of the overlapping areas (for polygons) or the distance (for points) represent the geographical similarity (*GS*) between the geometries of a pair.

The geometry pairs are categorized according to the classification rule shown in Table 1 which considers the similarity values as well as a linguistic similarity threshold (*LST*) and a geographic similarity threshold (*GST*). The geometry pairs are classified into threes categories: *match*, *non-match*, and *potential match*. A *match* indicates that the geometry pair is a correspondence, while *non-match* indicates no (or low) similarity between the geometries of the pair. A *potential match* indicates that the geometries present a high similarity according to only one matcher (i.e., linguistic or geographic).

The Entity Matching process supports the Data Quality assessment by providing information about the involved data sources. More precisely, the EM approach provides additional information about the data source reliability by identifying duplicate entities and exploiting correspondences between entities from different data sources. The aim is to analyze the amount of matching, non-matching and potential matching entities available in the data sources.

[1] https://github.com/eubr-bigsea/EMaaS

Table 1: Classification rules.

Rule	Classification
LS >*LST* and *GS* >*GST*	*Match*
LS <*LST* and *GS* <*GST*	*Non-Match*
(*LS* >*LST* and *GS* <*GST*) or (*LS* <*LST* and *GS* >*GST*)	*Potential match*

In this sense, the result of the EM approach over a single data source can be used to calculate a DQ dimension known as *Uniqueness*, i.e., the degree of duplicated entities in the data source. Each object should be represented by an unique entity, otherwise the risk of accessing erroneous information increases. Thus, let us consider $Card(S)$ as the cardinality of a data source, i.e., the number of entities represented in the data source and $Dupl(S)$ the number of duplicated entities retrieved by the EM approach, the Uniqueness (U) dimension can be defined as:

$$U(S) = \frac{Card(S) - Dupl(S)}{Card(S)} \times 100$$

3.2 Two data sources

Regarding the context in which a pair of data sources (R and S) is provided to the EM approach, the workflow is similar to the single data source approach. Each geometry (entity) of R is paired and compared with all geometries of S (Cartesian product). At each comparison between two geometries, the linguistic and geographical similarity values are determined by the linguistic and geographic matchers, respectively. Finally, the geometry pairs are classified as *match*, *non-math* or *potential match*, according to the classification rule presented in Table 1.

In order to support applications/users be aware of the integration quality involving two data sources, it is important to highlight three scenarios regarding the data sources where the Entity Matching can be applied: a) clean-clean, b) clean-dirty, and c) dirty-dirty. A data source is considered clean if U(S) = 100%. On the other hand, a data source is considered dirty if $U(S) <= \Phi_{dirty}$, where Φ_{dirty} is a threshold that can be defined according to the characteristics of the data source. For instance, for a data source in which a certain degree of dirtiness is negligible we can assume a Φ_{dirty} = 95%.

The EM execution over two data sources can be helpful to better understand the completeness of the data sources. In fact, considering two data sources R and S the completeness calculated on the single source R can be improved by considering the difference between the number of objects represented in the two sources. However, since the evaluation of the completeness can be compromised by the presence of dirty data (due to the linkage of different representations of the same object), it is important to remove all the duplicates from each single source before the evaluation of the completeness. For this, let the set S_{clean} (i.e., Clean S) be defined as $S - S_{duplicates}$ and the set R_{clean} (i.e., Clean R) be defined as $R - R_{duplicates}$, where $S_{duplicates}$ and $R_{duplicates}$ are the sets of

duplicated entities from data sources S and R, respectively. Thus, the completeness of S regarding R can be calculated as:

$$C(S,R) = \frac{Card(S_{clean})}{Card(S_{clean}) + Card(R_{clean}) - Card(S_{clean} \bowtie R_{clean})}$$

4 VALIDATION

In this section, we evaluate the EM approach using real-world data sources provided by Open Street Map[2] and the municipality of Curitiba (i.e., IPPUC). The evaluation addresses quantitative and qualitative issues in order to calculate the Uniqueness and Completeness data quality dimension of each data source. Furthermore, we analyze possible inconsistencies regarding the matching of geographic points and polygons contained in the data sources.

4.1 Matching of Geographical Points

Concerning a single data source, the EM approach analyzes the amount of duplicated entities. We analyze two data sources: *bus stops osm* (containing 736 entities) and *bus stops municipality* (containing 6,982 entities). Analyzing both data sources, we identified that they have only one attribute in common (coordinate). For this reason, we only consider a geographic matcher to perform the EM task. After performing the EM task over each data source, the results have shown that both data sources (*bus stops osm*, and *bus stops municipality*) do not present duplicated entities. Thus, they are considered clean data sources, i.e., their *uniqueness* is $U = 100$.

To integrate these data sources, the geographic matcher considers as *match* the entity pairs that have a distance (in meters) lower than a certain *GST* (also given in meters). Otherwise, the entity pair is considered as *non-match*. However, if we consider only the geographic distance, two problems can be highlighted. Since several bus stops may be close to each other in the real world, a bus stop can be classified as match more than once with different representations of bus stops from the other data source. Moreover, the application of the geographic matcher can only reduce the reliability of the results since the bus stops may be close but might not represent the same bus stop.

Figure 3 shows the amount of identified matches (left axis) and the completeness values (right axis) at the proportion that the distance, i.e, the matching threshold, between the bus stops increases. Regarding the identified matches, the increasing of the distance allows more geometry pairs to be classified as a match. However, this increasing also allows particular entities to be matched more than once. In this sense, the entities matched multiple times represent a problem, since a bus stop cannot correspond to more than one bus stop in the real world. Thus, we evaluate the completeness dimension for the following data sources: *bus stops osm* over *bus stops municipality* and *bus stops municipality* over *bus stops osm*. In Figure 3, we vary the distance (that consider as matches two bus stops) from 0 to 20 meters. Note that, as long as we increase the distance, the number of entities matched more than once with the entities belonging to the other data source is also increasing.

[2]http://www.openstreetmap.org

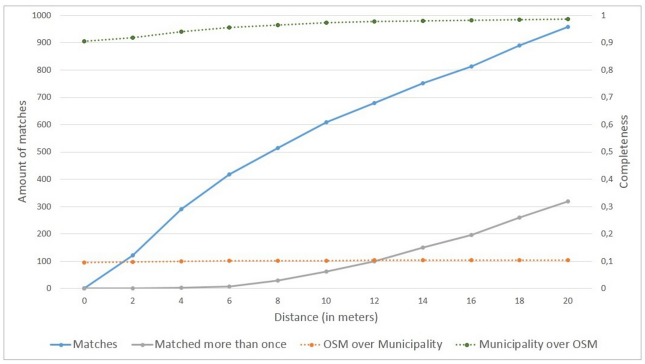

Figure 3: The EM results for *bus stops osm* and *bus stops municipality* data sources.

4.2 Matching of Geographical Polygons

Regarding a single data source, the EM approach analyzes the amount of duplicated entities. For instance, we analyze two data sources: *squares municipality* (containing 682 entities) and *squares osm* (containing 3,185 entities). The data source *squares municipality* does not present duplicated entities: it is a clean data source and its *uniqueness* is $U(squares\ municipality) = 100$. Data source *squares osm* contains instead 12 duplicated entities and thus its uniqueness $U(squares\ osm) = 99.6$.

In order to better understand the quality of the sources *squares municipality* and *squares osm*, the EM approach identifies similar or potential similar entities between the data sources. To this aim, we analyze the entity pairs considered as matches or possible matches, and we identify that 341 squares contained in *squares municipality* are identified as a match in the data source *squares osm*. Particularly, 340 entity pairs are considered a possible match, i.e., these pairs present a high similarity according to at least one match (linguistic or geographic). Based on the number of matches between both data sources and the number of duplicate pairs in each data source, the completeness of *squares municipality* regarding *squares osm* is $C(squares\ municipality, squares\ osm) = \frac{(682-0)}{((682-0)+(3185-12)-(341))} = 19.4$. On the other hand, the completeness of *squares osm* regarding *squares municipality* is $C(squares\ osm, squares\ municipality) = \frac{(3185-12)}{((3185-12)+(682-0)-(341))} = 90.3$. Uniqueness and completeness enrich the set of data quality dimensions for the analyzed data sources. The obtained quality values can support the identification of the suitable source to obtain valuable results. In the example, the completeness for *squares osm* is significantly higher than *squares municipality* and we can suggest this as the best source for further analysis.

5 RELATED WORK

In the literature, several papers claim that data quality dimensions need to be redefined [2, 5, 12]. For example, [2] focuses on the evolution of data quality dimensions and shows how the definitions of these dimensions change on the basis of data type, sources and applications considered. For example, they consider seven clusters of dimensions (i.e., accuracy, completeness, redundancy, readability, accessibility, consistency, and trust) and compare their definition

in structured databases to their definition when data of a different type (i.e., geographic data, linked data, and semi-structured text) are considered. In [5], the authors define that the quality dimensions, in quality assessment process, depends on the goals of data collection and thus also on the considered business environment and the involved data sources. A model that can be used to assess the level of quality-in-use of the data is proposed in [12]. Authors define the concept of Adequacy of data as "the state or ability of data of being good enough to fulfil the goals and purposes of the analysis". All these papers confirm the motivations behind our work: data quality dimensions definition and assessment algorithm have to be redefined since they are strongly dependent on the type of data and on the application that requests data. In this sense, our work proposes an architecture for assessing DQ in which the EM task is used to provide quality metrics (uniqueness and completeness) that can assist the Data Quality assessment.

Concerning the matching of geographic data, the survey [18] classifies and describes several works which propose matching approaches in the context of geographic data. The work [7] presents deduplication techniques based on a language model that can encapsulate both domain knowledge as well as local geographical knowledge. This model allows exploiting the domain knowledge and the spatial attributes to identify duplicated data. In [8], a polygon-based approach is proposed to match roads and urban blocks provided by Open Street Map with official data sources. The algorithm extracts urban blocks that are central elements of urban planning and are represented by polygons surrounded by their surrounding streets, and it then assigns road lines to edges of urban blocks by checking their topologies. Thus, the polygons of urban blocks can be compared, checking for overlapping areas, in order to determine urban blocks correspondents. Our work is also based on the overlapping areas between polygons to determine correspondent polygon pairs. In this terms, although the works previously discussed address the geographical data matching, none of them proposed approaches directly related to Data Quality assessment. Therefore, our work emerges as a bridge between Data Matching and Data Quality assessment.

Regarding the Entity Matching task, its importance is evident in Data Quality Assessment [15]. To make practical use of the data, it is often necessary to accurately identify entities from different sources that refer to the same object [9]. A variety of approaches have been proposed for entity matching and many of them are described in the recent surveys [16][6]. The main types of the approaches are probabilistic, learning-based, distance-based, and rule-based. In our work, we focus on rule-based Entity Matching in order to evaluate the Uniqueness and Completeness dimension to support Data Quality assessment.

6 CONCLUSIONS

In this paper, we propose an approach for assessing the quality of data sources contributing to smart city scenarios. The solution is able to perform the analysis of static data sources through the application of EM task. Based on the evaluation results, we have observed that i) the data quality metrics are essential to understand the fit for use of a source; and ii) the EM task can support data quality analysis in order to provide information about the level of duplicated entities available in a data source. Furthermore, the EM task can be used to assist data integration, since it identifies similar or potential similar entities between two data sources.

In future work, we intend to propose data integration using different data sources and a methodology to assess data quality in such a scenario. Regarding the EM task, we also intend to propose parallel EM approaches capable of scaling efficiently the EM task in the context of Big Data [1, 14], making it suitable also for streaming data sources. Moreover, we will analyze how our data quality approach can be applied to data sources from other smart city scenarios.

7 ACKNOWLEDGMENTS

We would like to thank the Municipality of Curitiba, IPPUC, CAPES, and CNPq. This work is partially funded by the National Science Foundation (NSF) grant HRD-1242122 and EU-BR EUBra-BigSea project (MCTI/RNP 3^{rd} Coordinated Call).

REFERENCES

[1] T. B. Araújo, C. E. S. Pires, T. P. da Nóbrega, and D. C. Nascimento. A fine-grained load balancing technique for improving partition-parallel-based ontology matching approaches. *Knowledge-Based Systems*, 111:17 – 26, 2016.
[2] C. Batini, A. Rula, M. Scannapieco, and G. Viscusi. From data quality to big data quality. *J. Database Manag.*, 26(1):60–82, 2015.
[3] C. Batini and M. Scannapieco. *Data and Information Quality - Dimensions, Principles and Techniques.* Data-Centric Systems and Applications. Springer, 2016.
[4] L. Berti-Équille and J. Borge-Holthoefer. *Veracity of Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics.* Synthesis Lectures on Data Management. Morgan & Claypool Publishers, 2015.
[5] L. Cai and Y. Zhu. The challenges of data quality and data quality assessment in the big data era. *Data Science Journal*, 14:2, 2016.
[6] P. Christen. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection.* Springer Science & Business Media, 2012.
[7] N. Dalvi, M. Olteanu, M. Raghavan, and P. Bohannon. Deduplicating a places database. In *Proceedings of the 23rd international conference on World wide web*, pages 409–418. ACM, 2014.
[8] H. Fan, B. Yang, A. Zipf, and A. Rousell. A polygon-based approach for matching openstreetmap road networks with regional transit authority data. *International Journal of Geographical Information Science*, 30(4):748–764, 2016.
[9] W. Fan. Data quality: From theory to practice. *SIGMOD Rec.*, 44(3):7–18, Dec. 2015.
[10] L. Kolb, A. Thor, and E. Rahm. Load balancing for mapreduce-based entity resolution. In *Proceedings of ICDE12*, pages 618–629, Washington, DC, USA, 2012. IEEE Computer Society.
[11] H. Kopcke and E. Rahm. Frameworks for entity matching: A comparison. *Data Knowl. Eng.*, 69(2):197–210, Feb. 2010.
[12] J. Merino, I. Caballero, B. Rivas, M. A. Serrano, and M. Piattini. A data quality in use model for big data. *Future Generation Comp. Syst.*, 63:123–130, 2016.
[13] D. G. Mestre and C. E. Pires. Efficient entity matching over multiple data sources with mapreduce. *Journal of Information and Data Management*, 5(1):40, 2014.
[14] D. G. Mestre, C. E. S. Pires, and D. C. Nascimento. Towards the efficient parallelization of multi-pass adaptive blocking for entity matching. *Journal of Parallel and Distributed Computing*, 101:27 – 40, 2017.
[15] F. Naumann. Data profiling revisited. *SIGMOD Rec.*, 42(4):40–49, Feb. 2014.
[16] F. Naumann and M. Herschel. *An Introduction to Duplicate Detection.* Morgan and Claypool Publishers, 2010.
[17] R. Y. Wang and D. M. Strong. Beyond accuracy: What data quality means to data consumers. *J. Manage. Inf. Syst.*, 12(4):5–33, Mar. 1996.
[18] E. Xavier, F. J. Ariza-López, and M. A. Ureña-Cámara. A survey of measures and methods for matching geospatial vector datasets. *ACM Computing Surveys (CSUR)*, 49(2):39, 2016.