

Assessing the Quality of Spatio-Textual Datasets in the Absence of Ground Truth

Mouzhi Ge^{1(✉)} and Theodoros Chondrogiannis^{2(✉)}

¹ Masaryk University, Brno, Czech Republic
`mouzhi.ge@muni.cz`

² Free University of Bozen-Bolzano, South Tyrol, Italy
`tchond@inf.unibz.it`

Abstract. The increasing availability of enriched geospatial data has opened up a new domain and enables the development of more sophisticated location-based services and applications. However, this development has also given rise to various data quality problems as it is very hard to verify the data for all real-world entities contained in a dataset. In this paper, we propose ARCI, a relative quality indicator which exploits the vast availability of spatio-textual datasets, to indicate how confident a user can be in the correctness of a given dataset. ARCI operates in the absence of ground truth and aims at computing the relative quality of an input dataset by cross-referencing its entries among various similar datasets. We also present an algorithm for computing ARCI and we evaluate its performance in a preliminary experimental evaluation using real-world datasets.

Keywords: Spatio-textual data · Data quality · Relative quality

1 Introduction

The current trends in technology, such as smartphones and sensor networks, along with the proliferation of location-based social networks, such as Foursquare and Flickr, have motivated the development of new applications and services which employ spatio-textual datasets, i.e., collections of spatial objects which carry both spatial and textual information. In addition, spatio-textual queries, which combine location-based retrieval with textual information that describes the spatial objects, have recently attracted much attention [3, 15].

The fact that real world entities are constantly changing though, e.g., restaurants are closing down, museums are being moved etc., makes the maintenance of spatio-textual datasets very hard. Consequently, it is quite common for spatio-textual datasets to contain inaccurate information and/or incomplete entries [7]. Furthermore, data across different datasets is not always consistent, i.e., entries of different datasets referring to the same real-world entity may provide conflicting information. Due to such data quality issues, a number of geospatial applications and initiatives have been delayed or even canceled, citing poor-quality

of the available data as the main reason. Although identifying such data quality problems manually is unrealistic, it is important to provide at least an indication for the quality of spatio-textual datasets. Such indicators should operate in the absence of ground truth, i.e., when there is no available verified data.

In this paper, we propose ARCI, a novel approach to indicate the relative quality of a spatio-textual dataset by cross-referencing its entries with other spatio-textual datasets. The ARCI indicator intends to highlight the trustfulness of the dataset and confirm the confidence to use the data. Furthermore, quantifying the ARCI indicator allows us to visualize the data quality, for example using a spectrum bar in this paper. While most existing quality assessment methodologies [6, 7] rely on (designed) ground truth data, our approach works in the absence of ground truth. Instead of focusing on a specific data quality dimension such as accuracy or completeness, ARCI uses data alignment as a quality indicator for the input dataset. We also present ARCI-GP, an algorithm for computing ARCI and we conduct a preliminary experimental evaluation.

The remainder of the paper is organized as follows. Section 2 overviews the related work. Section 3 presents some preliminaries on spatio-textual datasets, spatio-textual similarity search and spatio-textual similarity joins. Our approach for evaluating the relative quality of spatio-textual data is introduced in Sect. 4. Section 5 reports on the results of a preliminary experimental evaluation. Finally, Sect. 6 concludes the paper and outlines future work.

2 Related Work

A data value can be considered dirty if it does not conform to the reference data [5]. However, in reality it is usually unrealistic to define a set of comprehensive and correct reference data as the ground truth. Thus, relative data quality has been recently emerging as a research trend. Cao et al. [4] have defined the relative accuracy of data entries as the closeness of the value of different entries, and further specified that the challenge in relative accuracy is creating accuracy rules and inferring the true value. Their work is still built upon the availability of master data though.

Regarding the absence of ground-truth data, Galarus and Angryk [7] have further elaborated this challenge for cases where it is not feasible to obtain the ground-truth data at all times. In order to leverage this issue, they have developed a representative artificial dataset, which can be used as an interpolator to estimate unknown values. The focus of their approach is to mitigate the erroneous data that can possibly appear during the data processing. In their context, it is possible to construct the representative dataset from historical data, which as facts will not be changed.

Another way to detect data quality problems without using reference data is to define data quality rules. In order to discover such rules, different approaches have been proposed such as conditional functional dependencies [5], association rule learning [1] or defining quality processing requirements from users [11]. These rules are usually determined or inferred in certain context. In this work,

the experiment has been conducted context-independently and the datasets used contain a limited set of attributes. Therefore, it is not suitable for this work to use rules to detect data quality problems.

Furthermore, data quality problems are usually related to certain data quality dimensions. The research of data quality dimensions can be traced back to the 90's. Wang and Strong [16] used an exploratory factor analysis to derive fifteen data quality dimensions, which are widely accepted in the subsequent data quality research. Afterwards, different data quality dimensions have been further studied and refined such as consistency, accuracy [6] or completeness [13]. This paper does not contribute to a specific data quality dimension, but instead introduces an overall quality indicator to show the trust level of the data.

3 Spatio-Textual Similarity Search

Building upon the definition of spatio-textual objects in [3], a spatio-textual dataset $D = \{x_1, \dots, x_k\}$ is a collection of spatio-textual objects $x_i = \langle x.id, x.txt, x.loc, x.attr \rangle$ where $x.id$ is the id of the object within the dataset, $x.txt$ is a textual attribute of x (i.e. the name), $x.loc$ is the location of x in the two dimensional geographical space and $x.attr$ is the attribute which is to be assessed.

To determine whether two spatio-textual objects refer to the same real-world entity, one can aim either for exact or for approximate matching. Given two spatio-textual objects x_1 and x_2 , for exact matching $x_1.txt$ must be equal to $x_2.txt$, and $x_1.loc$ must also be equal to $x_2.loc$. Clearly, such an approach is not practical as tiny differences, especially in the spatial attribute (i.e. the locations of the objects are few centimeters apart) will prevent a matching. For approximate matching, we employ a similarity function $sim(x_1.txt, x_2.txt)$ for the textual attribute and a distance function $dist(x_1.loc, x_2.loc)$ for the spatial attribute. Hence we declare that x_1 and x_2 refer to the same real-world entity if $sim(x_1.txt, x_2.txt) > \theta$ and $dist(x_1.loc, x_2.loc) \leq \epsilon$, where θ is a textual similarity threshold, and ϵ is the maximum distance that two locations can be apart. Since attributes $x.txt$ and $x.loc$ are used to identify one spatio-textual object from another, throughout this paper, we refer to these two attributes as the identifiers of x .

To perform the spatio-textual similarity search, first we need to decide on a similarity metric for textual attributes, and a distance metric for spatial attributes. For textual similarity in particular, a variety of metrics has been proposed [14]. In this paper, we employ the Levenshtein [10] distance to compute the spatio-textual similarity and we also run experiments using the N-Gram textual similarity [8]. However, choosing the most suitable similarity metric is out of the scope of this paper. Regarding the spatial distance between two objects, it is given by their geodetic distance.

3.1 Spatio-Textual Similarity Joins

An efficient approach to achieve approximate matching between objects from different datasets is to perform a *Spatio-textual similarity join* (STJoin) [2, 3, 12]. STJoin queries aim at identifying similar spatio-textual objects across different datasets that are close to each other and have similar textual descriptions. More specifically, given two datasets D_i and D_j , a textual similarity threshold ϵ and a spatial distance threshold θ , a spatio-textual similarity join query retrieves all pairs of objects $\{x_i, x_j\}$ where $x_i \in D_i$ and $x_j \in D_j$, the spatial distance of x_i and x_j is $\text{dist}(x_i.\text{loc}, x_j.\text{loc}) \leq \theta$ and the textual similarity is $\text{sim}(x_i.\text{txt}, x_j.\text{txt}) > \epsilon$.

Various algorithms for processing STJoin queries have been proposed. According to the results of the experimental evaluation in [3], the most efficient algorithm for evaluating STJoin queries is PPJ-C algorithm. Given a spatio-textual dataset, PPJ-C first defines a dynamic grid partitioning such that the length of the side of each grid cell is equal to the distance threshold ϵ . Next, for each cell C , PPJ-C identifies the set of join cells A_C , i.e., cell C along with its adjacent cells. Finally, for each pair of join cells, the algorithm compares the elements in the two cells and adds to the result set all pairs of objects that satisfy both the textual similarity and the spatial distance constraint.

4 Spatio-Textual Data Quality Assessment

In this section, we describe our methodology for assessing the relative quality of a spatio-textual dataset. Let a spatio-textual dataset D be the dataset the quality of which we will evaluate. Since we operate in the absence of ground truth, the main idea behind our approach is to cross-reference the entries of D with other similar spatio-textual datasets. For example, given a dataset D_e the quality of which we want to assess, and another dataset D_r , we first match each spatio-textual object $x \in D_e$ with some spatio-textual object $y \in D_r$ such that x and y refer to the same real-world entity. Then, for each attribute of x we compute a relative correctness indicator by comparing the attributes of x with the attributes of y .

4.1 Attribute Relative Correctness Indicator

We propose the *Attribute Relative Correctness Indicator* (ARCI) to assess the quality of a spatio-textual dataset. Let D_e be the spatio-textual dataset the quality of which we want to assess and D_r be another spatio-textual dataset. Each dataset contains spatio-textual objects $x = \langle x.\text{id}, x.\text{txt}, x.\text{loc}, x.\text{attr} \rangle$, where $x.\text{txt}$ and $x.\text{loc}$ are the identifiers of x . Given a textual similarity threshold θ , and a spatial distance threshold ϵ , for every spatio-textual object $x_i \in D_e$ we compute the spatio-textual object $y_j \in D_r$ which is the most similar to x_i and does not violate the θ and ϵ thresholds (it is possible that no matching spatial-object is found). Having matched all possible spatial objects, the ARCI of $x_i.\text{attr}$ is

$$ARCI(x_i.attr, y_j.attr) = \begin{cases} Sim(x_i.attr, y_j.attr), & \text{if } y_j \text{ is the best match to } x_i \\ 0.5, & \text{if } \nexists y_j \text{ that matches } x_i. \end{cases}$$

Figure 1 illustrates two datasets: Dataset 1 (D_e), the dataset the quality of which we evaluate, and Dataset 2, (D_r) another dataset we use as reference. The identifiers of both D_e and D_r are the *name* and the *location*, while the attribute for which we want to compute the ARCI is the *type*. The textual similarity is given by the Levenshtein distance, while the spatial distance is given by the Euclidean distance. Finally, we set the textual similarity threshold $\epsilon = 0.6$ and the spatial distance threshold $\theta = 1$.

Dataset 1 (D_{eval})				Dataset 2 (D_{ref})			
id	name	type	loc.	id	name	type	loc.
x_1	Loacker	Café	(1, 1)	y_1	Da Pichio	Bar	(4, 5)
x_2	Marilyn	Café	(3, 2)	y_2	Loacker	Café	(1, 1)
x_3	Cavalino Bianco	Restaurant	(1, 6)	y_3	Nadamas	Bar	(5, 1)
x_4	Dai Carrettai	Restaurant	(5, 5)	y_4	Marylin	Café	(3, 3)
x_5	Enovit Bar	Bar	(8, 9)	y_5	Enovit Wine Bar	Wine Bar	(9, 8)
x_6	Hopfen	Brewery	(7, 5)	y_6	Stadt Café Città	Café	(8, 2)
				y_7	Nussbaumer	Restaurant	(3, 7)
				y_8	Hopfen & Co.	Restaurant	(7, 5)
				y_9	Carrettai	Restaurant	(7, 7)

Fig. 1. Sample spatio-textual datasets.

First, we attempt to match each entry $x_i \in D_e$ with an entry $y_i \in D_r$. The result of this operation is illustrated in Fig. 2. Having computed the matching spatio-textual objects, we compute the ARCI for the attribute "type" for each spatio-textual object in D_e . For x_1 and x_2 we have an exact match, hence $ARCI(x_1.attr) = 1$ and $ARCI(x_2.attr) = 1$. For x_1 and x_2 the ARCI indicates that both $x_1.attr$ and $x_2.attr$ are correct with regard to D_r . For, x_3 and x_4 no matching object in D_r was found; thus $ARCI(x_3.attr) = 0.5$ and $ARCI(x_4.attr) = 0.5$ meaning that it was not possible to verify the correctness of the attributes. Finally, for elements x_5 and x_6 the respective ARCIs are $ARCI(x_5.attr) = 0.36$ and $ARCI(x_6.attr) = 0.1$ which are relatively low and, therefore, our approach indicates that $x_5.attr$ and $x_6.attr$ might be wrong.

$x_i \in D_{eval}$	$y_j \in D_{ref}$	$Sim(x_i.txt, y_j.txt)$	$dist(x_i.loc, y_i.loc)$	ARCI
x_1	y_2	1	0	1
x_2	y_4	0	1	1
x_3	-	-	-	0.5
x_4	-	-	-	0.5
x_5	y_5	0.667	1.414	0.36
x_6	y_8	0.667	0	0.1




Fig. 2. Matching entries of D_e and D_r of Fig. 1.

Note that ARCI cannot be computed for the identifiers of a spatio-textual object. Apparently, the selection of the identifiers (especially the textual one) is crucial and can possibly affect the efficiency of our approach. However, as determining the best possible identifiers of a spatio-textual object is out of the scope of this paper, we work under the assumption that each spatio-textual object comes with the proper identifiers.

4.2 The ARCI-GP Algorithm

To compute the ARCI for each spatio-textual object in a dataset we propose **ARCI-GP**, an algorithm which is inspired by the **PPJ-C** [3] algorithm for spatio-textual similarity joins. Let D_e be the spatio-textual dataset the quality of which we want to assess and D_r be the reference dataset. First, **ARCI-GP** defines a dynamic grid partitioning and organizes the spatio-textual objects of both D_e and D_r into cells. The grid is defined such that the length of the side of every cell equals the spatial distance threshold ϵ . For each cell C , **ARCI-GP** identifies a set of join cells A_C , i.e., the set of the adjacent cells of C along with C itself. For every element $x \in C$ of D_e , A_C contains the objects $y \in D_r$ which can possibly match with x . Due to the properties of the grid, the distance of all other elements $y' \notin A_C$ from x is more than ϵ , i.e., they violate the spatial distance constraint. Next, **ARCI-GP** computes for each object $x \in C$ of D_e the object $y \in A_C$ of D_r which is the most similar to x . To determine the best match of x , the algorithm examines only the objects in A_C . During this process, it is possible that the algorithm does not find any matching object to x . The same process is executed over all the cells that contain at least one object $x \in D_e$.

Algorithm 1 illustrates the pseudocode of our **ARCI-GP** algorithm. First, the result set R is initialized in Line 1 and a grid partition G_R is constructed for the spatio-textual objects in both input datasets D_e and D_r . In Lines 3–16 the algorithm iterates over the cells of the partition that contain at least one element of D_e and, for each cell C , it computes the set of join cells A_C . Then, for each cell $C' \in A_C$, the algorithm attempts to find a match between the element $o_i \in C$ that is an element of D_e , and the element $o_j \in C'$ that is an element of D_r (Lines 6–16). A new entry r is initialized to *null* in Line 7. Each element o_i is matched with only one element o_j for which the textual similarity $Sim(o_i.txt, o_j.txt)$ is maximum and the spatial distance $dist(o_i.loc, o_j.loc)$ is minimum (Lines 8–12). If a match is found, i.e. r is not *null*, then r is added to the result set with the computed ARCI value in Line 14. Otherwise, r is added to the result set with the default 0.5 ARCI value. Finally, the result set is returned in Line 17.

Algorithm 1. ARCI-GP ($D_e, D_r, \epsilon, \theta$)

Input: Collection of spatio-textual objects D_e ; collection of spatio-textual objects D_r ; spatial distance threshold ϵ ; textual similarity threshold θ

Output: Result set R

```

1  $R \leftarrow \emptyset$ ;
2  $G_R \leftarrow \text{ConstructGridPartitioning}(D_e \cup D_r, \epsilon)$ ;
3 foreach cell  $C \in G_R$  do
4    $A_C \leftarrow \text{IdentifyJoinCells}(G_R, C)$ ;
5   foreach cell  $C' \in A_C$  do
6     foreach object  $o_i \in C \cap D_e$  do
7       initialize entry  $r \leftarrow \text{null}$ ;
8        $\text{sim}_{\max} = \theta$ ;
9       foreach object  $o_j \in C \cap D_r$  do
10        if  $\text{Sim}(o_i.\text{txt}, o_j.\text{txt}) > \text{sim}_{\max}$  and  $\text{dist}(o_i.\text{loc}, o_j.\text{loc}) \leq \epsilon$  then
11           $r \leftarrow \langle o_i, \text{ARCI}(o_i, o_j) \rangle$ ;
12           $\text{sim}_{\max} \leftarrow \text{Sim}(o_i.\text{txt}, o_j.\text{txt})$ ;
13        if  $r \neq \text{null}$  then
14           $R \leftarrow R \cup r$ ; ▷ Best match added to result
15        else
16           $R \leftarrow R \cup \langle o, 0.5 \rangle$ ; ▷ No match was found
17 return  $R$ ;
```

5 Preliminary Experimental Evaluation

In this section, we report the results of a preliminary experimental evaluation and compare two different implementations of our ARCI-GP algorithm. We use two different spatio-textual datasets in our experiments. The first dataset (D_e) contains 500,000 spatio-textual objects and was compiled by combining datasets obtained from Tourpedia¹. The second dataset (D_r) contains 1,000,000 spatio-textual objects and was obtained by querying the Foursquare API² using *ids* obtained from [9]. To observe the effect of the textual similarity metric, we measure the runtime of ARCI-GP using two different metrics: the Levenshtein similarity [10] and the NGram similarity [8]. We implemented our algorithm with Java 1.8 and the tests run on a machine with 4 Intel Xeon X5550 (2.67 GHz) processors and 48 GB main memory running Ubuntu Linux.

Figure 3 shows our measurement results on the runtime of our ARCI-GP algorithm by varying the sizes of D_e and D_r . More specifically, Fig. 3a shows the runtime of ARCI-GP varying the size of D_e from 100,000 to 500,000 entries using the entire D_r dataset, and Fig. 3b shows the runtime of ARCI-GP varying the size of D_r from 200,000 to 1,000,000 entries using the entire D_r dataset. In both figures we observe that the runtime of ARCI-GP increases with the size of

¹ <http://tour-pedia.org/about/datasets.html>.

² <https://developer.foursquare.com>.

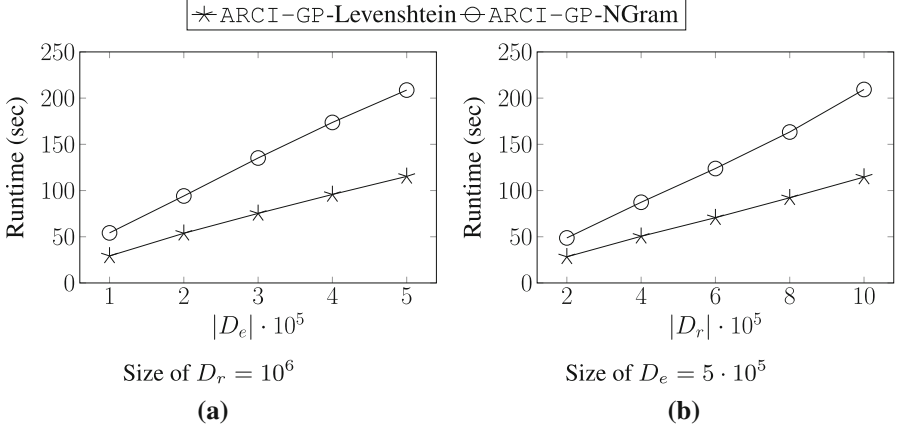


Fig. 3. Performance of PPLJ.

D_e and D_r . However, we can observe that the runtime increases faster with the size of D_e . For example, the runtime for $D_e = 2 \cdot 10^5$ and $D_r = 10^6$ (Fig. 3a) and the runtime for $D_e = 4 \cdot 10^5$ and $D_r = 5 \cdot 10^5$ (Fig. 3b). Although in the first case the total number of involved entries is higher by 300,000 entries, the runtime is approximately the same. Apparently, the size of D_e influences the runtime of ARCI-GP much more than the size of D_r . Such a result is to be expected as ARCI-GP considers all the elements in D_e , while many of the elements in D_r are filtered out by the grid partitioning.

Finally, with regard to the textual similarity metric, we observe that our algorithm is almost two time faster when using the Levenshtein similarity than when using the NGram similarity. Since the algorithm requires the execution of many textual similarity operations, the computational cost of the similarity metric has a great influence on the total runtime of the ARCI-GP.

6 Conclusion

In this paper, we proposed ARCI, an indicator which operates in the absence of ground truth and shows the relative quality of a spatio-textual dataset by cross-referencing it with similar datasets. ARCI is computed for the attributes of a data entry in a spatio-textual dataset to indicate its relative correctness. We have also shown that ARCI can be used directly to provide visual information on the quality of the data, e.g., using a spectrum bar. Furthermore, we proposed an algorithm for computing ARCI and evaluated its performance using real-world spatio-textual datasets.

As future work, we will explore different strategies to develop efficient algorithm for computing ARCI such as executing textual-matches first. We will also consider utilizing indexing structures to improve performance even further. Finally, we plan to investigate alternative metrics for computing textual similarity such as semantic similarity metrics.

References

1. Abedjan, Z., Akcora, C.G., Ouzzani, M., Papotti, P., Stonebraker, M.: Temporal rules discovery for web data cleaning. *Proc. VLDB Endowment* **9**(4), 336–347 (2015)
2. Ballesteros, J., Cary, A., Rishe, N.: Spsjoin: parallel spatial similarity joins. In: *Proceedings of the 19th ACM SIGSPATIAL GIS Conference*, pp. 481–484 (2011)
3. Bouros, P., Ge, S., Mamoulis, N.: Spatio-textual similarity joins. *Proc. VLDB Endowment* **6**(1), 1–12 (2012)
4. Cao, Y., Fan, W., Yu, W.: Determining the relative accuracy of attributes. In: *Proceedings of the 2013 ACM SIGMOD Conference*, pp. 565–576 (2013)
5. Chiang, F., Miller, R.J.: Discovering data quality rules. *Proc. VLDB Endowment* **1**(1), 1166–1177 (2008)
6. Cong, G., Fan, W., Geerts, F., Jia, X., Ma, S.: Improving data quality: consistency and accuracy. In: *Proceedings of the 33rd VLDB Conference*, pp. 315–326 (2007)
7. Galarus, D., Angryk, R.: A smart approach to quality assessment of site-based spatio-temporal data. In: *Proceedings of the 24th ACM SIGSPATIAL GIS Conference*, pp. 55:1–55:4 (2016)
8. Kondrak, G.: N-gram similarity and distance. In: Consens, M., Navarro, G. (eds.) *SPIRE 2005. LNCS*, vol. 3772, pp. 115–126. Springer, Heidelberg (2005). doi:[10.1007/11575832_13](https://doi.org/10.1007/11575832_13)
9. Levandoski, J.J., Sarwat, M., Eldawy, A., Mokbel, M.F.: Lars: a location-aware recommender system. In: *Proceedings of the 28th IEEE ICDE*, pp. 450–461 (2012)
10. Levenshtein, V.: Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Phys. Doklady* **10**, 707–710 (1965)
11. Missier, P., Embury, S., Greenwood, M., Preece, A., Jin, B.: Quality views: capturing and exploiting the user perspective on data quality. In: *Proceedings of the 32nd VLDB Conference*, pp. 977–988 (2006)
12. Rao, J., Lin, J., Samet, H.: Partitioning strategies for spatio-textual similarity join. In: *Proceedings of the 3rd ACM International Workshop on Analytics for Big Geospatial Data*, pp. 40–49 (2014)
13. Razniewski, S., Nutt, W.: Completeness of queries over incomplete databases. *Proc. VLDB Endowment* **4**(11), 749–760 (2011)
14. Recchia, G., Louwerse, M.: A comparison of string similarity measures for toponym matching. In: *Proceedings of The 1st ACM International COMP Workshop*, pp. 54:54–54:61 (2013)
15. Tsatsanifos, G., Vlachou, A.: On processing top-k spatio-textual preference queries. In: *Proceedings of the 18th EDBT Conference*, pp. 433–444 (2015)
16. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. *J. Manage. Inf. Syst.* **12**(4), 5–33 (1996)

New Trends in Databases and Information Systems
ADBIS 2017 Short Papers and Workshops, AMSD,
BigNovelTI, DAS, SW4CH, DC, Nicosia, Cyprus,
September 24–27, 2017, Proceedings
Kirikova, M.; Nørnvåg, K.; Papadopoulos, G.A.; Gamper, J.;
Wrembel, R.; Darmont, J.; Rizzi, S. (Eds.)
2017, XVI, 434 p. 129 illus., Softcover
ISBN: 978-3-319-67161-1