Data quality: assessment and improvement of a dataset containing flight information

(Final project report of Data Architectures)

Claudio Rota, 816050
Simone Paolo Mottadelli, 820786

February 2020

Contents

1 Introduction

This project had the goal of assessing and improving a dataset containing data about the flight domain. First of all, the dataset has been standardized to make the data comparable. Afterwards, it has been assessed with respect to 3 data quality dimensions, that is, *completeness*, *consistency* and *duplication*. Then, the data quality has been improved by performing a record linkage phase, whose objective was to identify the records representing the same flight and group them into blocks. Finally, the data inside each block have been fused. These operations have led to the creation of a new dataset of a higher quality, which has been assessed another time, also considering the errors introduced by the improvement phases. The final outcome has demonstrated that the applied methodologies and techniques have effectively increased the aforementioned data quality dimensions.
As regards the implementation, this project has been realized with a script written in Java.
This project report is organized as follows: Section 2 gives an overview of the dataset used for this work, Section 3 explains how the data have been standardized, Section 4 describes the data quality dimensions studied and reports the assessment results before the improvement of the dataset, Section 5 explains how the data improvement process has been carried out and Section 6 concludes this report showing the results obtained after having assessed the data quality dimensions on the improved dataset.

2 Dataset description

The dataset (available at `http://lunadong.com/fusionDataSets.htm`) contained structured data about flights recorded in the period from 1 December 2011 to 3 January 2012 and coming from 38 sources: 3 airline websites, 8 airport websites and 27 third-party websites.
In particular, the dataset contained 776067 records and each of which had the following attributes: *source*, *flight code*, *scheduled departure*, *actual departure*, *departure gate*, *scheduled arrival*, *actual arrival* and *arrival gate*.
At first sight, the dataset appeared with a number of *null* and *pseudo-null* values, except for the *source* and the *flight code* attributes, which were also the only ones to be standardized. In fact, the other attributes seemed of a worse quality, as their values were represented in different formats and were often missing.
The authors of the dataset also provided a gold standard, which included 2986 high quality flights that were also contained in the dataset described above.

3 Dataset standardization

Initially, the data have been standardized in order to make the assessment and the improvement phases possible. The standardization was essential for the attributes of date-time type because their values were of heterogeneous formats, as every source has its way to represent such values: time was represented either in 12h or in 24h and the date was represented using different notations. In fact, there were at least 15 different data formats that had to be standardized to be comparable and Table 1 shows 6 different examples of date-time formats extracted from the dataset.
After this phase, all the values of date-time type followed this format:

month day year hour:minute timezone

All the departure and arrival dates have been standardized conforming to the time zone of the respective airports. The information about the time zone of each airport has been retrieved from the official *Bureau of Transportation Statistic* website (`https://www.bts.gov/`). Exceptions to this standard format have been made for the cancelled flights, which were not provided with a value for the time but just with a value for the date. In this particular case, the standard date format was the following:

month day year Cancelled

Concerning the *source* and the *flight code* attributes, their values appeared already standardized and, thus, no further operations were needed.
In addition, as regards the *departure gate* and the *arrival gate* attributes, despite they had different formats, it has not been possible to standardize their values in a reasonable time, because it would have been necessary to manually retrieve their correct format from their belonging airport websites, which were 249 in total.

Another standardization operation has been performed on the *pseudo-null* values, that is, values equivalent to the *null* value, such as *NO*, -, - -, *Not Available*, *Not provided by airline*, *HOLD* and *Contact the airline*. These values have been simply replaced with the *null* value.
Table 2 shows two records before and after the standardization phase. At that point, all the values had the same format and could be easily compared.

06:20 PM Thu 01-Dec-2011
2011-12-01 06:20 PM
6:20 PM, Dec 01
Dec 01 - 6:20 pm
Thu, Dec 1 6:20 PM
01/12/2011 18:20
12/1/2011 6:20 AM PST

Table 1: Some examples of date-time values

Source	Flight code	Scheduled departure	Actual departure	Departure gate	Scheduled arrival	Actual arrival	Arrival gate
flightexplorer	AA-1007-TPA-MIA	12/01/2011 01:55 PM	12/01/2011 02:07 PM	F78	12/01/2011 03:00 PM	12/01/2011 02:57 PM	D5
orbitz	UA-6436-LAX-ABQ	Tue, Dec 20 8:50 PM	Tue, Dec 20 8:50 PM	86	Tue, Dec 20 11:48 PM	Tue, Dec 20 11:48 PM	A3

Source	Flight code	Scheduled departure	Actual departure	Departure gate	Scheduled arrival	Actual arrival	Arrival gate
flightexplorer	AA-1007-TPA-MIA	Dec 01 2011 13:55 EST	Dec 01 2011 14:07 EST	F78	Dec 01 2011 15:00 EST	Dec 01 2011 14:57 EST	D5
orbitz	UA-6436-LAX-ABQ	Dec 20 2011 20:50 PST	Dec 20 2011 20:50 PST	86	Dec 20 2011 23:48 MST	Dec 20 2011 23:48 MST	A3

Table 2: Example of tuples before and after the standardization

4 Data quality assessment

The dataset has been evaluated with respect to 3 data quality dimensions:

- *completeness*: the degree to which a given data collection includes data describing the corresponding set of real-world objects [1];
- *consistency*: it refers to the violation of semantic rules defined over a set of data items [1];
- *duplication*: a measure of unwanted duplication existing within or across systems for a particular field, record, or data set [2].

4.1 Completeness

This dimension has been evaluated on every tuple, on every attribute and on the entire relation.
In order to measure this dimension, the following metric has been computed:

$$1 - \frac{\text{number of missing values}}{\text{total number of values}}$$

A value has been considered as missing if it was *null*. However, this definition had to be refined for the attributes of date-time type. The value of the *scheduled departure* and the *scheduled arrival* attributes had to contain both the date and the time to be considered as complete. On the other hand, since a flight could have been cancelled, the value of the *actual departure* and the *actual arrival* attributes had to contain either both the date and the time or the *Cancelled* keyword in order to be considered as complete. If these attributes did not respect the aforementioned definition, they would have been considered as incomplete.
First, the metric has been computed at the tuple level and the results in Table 3 show that only 23.94% of the total number of tuples were complete, 43.48% of them were half complete while 28.29% of the tuples had 2 missing attributes.
Then, the metric has been computed at the attribute level and Table 4 shows that all the records had a non-missing value for the *source* and the *flight code* attributes, while the *departure gate* and the *arrival*

gate were the ones with the highest number of missing values.
In conclusion, the relation completeness, which summarizes the previous results, was 70%. This means that almost a third of the data of the dataset was missing.

Number of tuples	Completeness	Missing values
35	37.5%	5
337445	50%	4
18462	62.5%	3
219566	75%	2
14767	87.5%	1
185792	100%	0

Table 3: Tuple completeness

Attribute	Completeness
source	100%
flight code	100%
scheduled departure	66.17%
actual departure	78.49%
departure gate	36.72%
scheduled arrival	66.37%
actual arrival	76.91%
arrival gate	35.77%

Table 4: Attribute completeness

4.2 Consistency

In this specific domain, different constraints have been identified. For instance, a flight cannot take off from an airport and land to the same airport, as well as the date and the time of the departure must precede the ones of the arrival. Table 5 shows the different constraints taken into account during the consistency assessment of the dataset. In particular, the constraint C2 imposes that a cancelled flight cannot have an actual departure time and an actual arrival time because it cannot be neither departed nor landed. Moreover, in the constraints C5 and C6, the minimum flight time has been computed in the following way:

$$\frac{\text{distance between A and B}}{\text{max speed for a service airplane}}$$

where A is the departure airport, B is the arrival airport and the maximum speed of a service airplane is about 575 mph. The information regarding the distance for every couple of airports has been retrieved from the same source also used for the time zones (`https://www.bts.gov/`).
Most of the tuples of the dataset appeared consistent with respect to the constraints previously defined. In fact, 6.53% of the records resulted inconsistent to the constraint C2, 2.03% to the constraints C3 and 1.58% to the constraint C4. As regards the other constraints, almost all the tuples resulted consistent. Table 6 shows 3 examples of inconsistent tuples, where the first one violated the C2 constraint, the second tuple the C3 constraint and the last tuple both the C3 and C4 constraints.

ID	Constraint	Consistent	Inconsistent
C1	The departure and the arrival airports must be different	100%	0%
C2	The cancelled flights must not have a time for the *actual departure* and the *actual arrival* attributes	93.47%	6.53%
C3	*scheduled departure* < *scheduled arrival*	97.97%	2.03%
C4	*actual departure* < *actual arrival*	98.42%	1.58%
C5	*scheduled departure* + min. flight time < *scheduled arrival*	99.98%	0.02%
C6	*actual departure* + min. flight time < *actual arrival*	99.93%	0.07%

Table 5: Consistency

Source	Flight code	Scheduled departure	Actual departure	Departure gate	Scheduled arrival	Actual arrival	Arrival gate
ua	UA-2703-CLT-CLE	Dec 01 2011 13:15 EST	Dec 01 2011 Cancelled	B11	Dec 01 2011 14:52 EST	Dec 01 2011 14:55 EST	null
flightstats	AA-185-JFK-LAX	Dec 01 2011 21:00 EST	Dec 01 2011 21:04 EST	null	Dec 01 2011 00:23 PST	Dec 01 2011 23:35 PST	null
weather	AA-185-JFK-LAX	Dec 01 2011 21:00 EST	Dec 01 2011 21:32 EST	null	Dec 01 2011 00:23 PST	Dec 01 2011 00:08 PST	null

Table 6: Examples of inconsistent tuples

4.3 Duplication

With reference to the duplication dimension, since a flight can be identified by its flight code and the departure date (time is excluded), two records represent the same flight if they have the same values for the aforementioned features. With this definition, if a flight had missing values for both the *scheduled departure* and the *actual departure* attributes, it was not identifiable, as a flight code is dependent to the departure date. In the same way, a flight without a flight code value was not identifiable, neither. For this reason, 5771 records out of 776067 (0.7%) have been excluded during the assessment of this dimension. This dimension has been evaluated by using the following metric:

$$1 - \frac{\text{number of distinct identifiable records}}{\text{total number of identifiable records}}$$

The computed value was 95.52%, which means that just a few flights were represented once in the dataset. This value is not surprising, as the dataset has been composed by integrating the data from different sources.

5 Data improvement

Soon afterwards having assessed the quality of the dataset considering the 3 data quality dimensions described in Section 4, different data quality improvement techniques have been applied to improve the dataset.

5.1 Record linkage phase

The record linkage phase aimed at identifying and grouping records referring to the same real world object into blocks, where two records correspond to the same flight if they had the same flight code and departure date, without considering the departure time. As already mentioned in Subsection 4.3, a record not provided with either a flight code or a departure date could not be identified, as the values of the other attributes were not sufficient to determine which flight a certain record referred to. Yet, it is important to underline that 258968 records out of 776067 (33%) had a missing value for the *scheduled departure* attribute, which was tightly dependent to the flight code value. On the other hand, removing 33% of the dataset was not a good idea. To address this problem, since the departure date extracted from the *scheduled departure* attribute was different from the one provided by the *actual departure* attribute in just 51 records out of 776067 (0.00007%), it has been possible to use the date value provided by the *actual departure* attribute when the one of the *scheduled departure* was missing, being aware that possible errors could have been introduced. However, as already mentioned in Subsection 4.3, 5771 records out of 776067 (0.7%) missed the value for both the *scheduled departure* and *actual departure* attributes and, thus, they have been removed also in this phase.
As a result of this phase, the remaining 770296 records have been grouped in 34812 blocks, each of which corresponded to a particular flight.

5.2 Data fusion phase

The blocks generated in the previous phase have been processed with the aim of fusing the records inside each block, obtaining a single record that could represent a certain flight. Such fusion phase has been carried out by considering all the records inside each block and by applying multiple conflict resolution and fusion strategies.
More specifically, since different records inside a specific block could have had missing or conflicting values, such as the time of the *scheduled departure* or the value of the *arrival gate*, it was impossible to establish for sure the true value of an attribute. For this reason, 3 conflict resolution strategies have been used: *cry with the wolves*, *trust your friends* and *take the information* [3]. The first strategy considers as correct value the most frequent one, the second strategy takes the value from the most trustworthy source, while the third one prefers concrete values over missing values. In this context, the most trustworthy sources are the airline websites, followed by the airport websites.
In general, to select the value of an attribute, first the *cry with the wolves* strategy has been applied, but, if either the most frequent value was *null* or there was not a majority for a specific value, then the *trust your friends* strategy has been alternatively applied. With the *trust your friends* strategy, the preferred values were the ones provided by the airline website, but, in case such values were missing, the ones of the airport websites have been selected instead. For the attributes regarding the departure, that is, the *scheduled departure*, the *actual departure* and the *departure gate*, the values provided by the departure airport website have been preferred over the ones of the arrival airport website, while for the attributes regarding the arrival, that is, the *scheduled arrival*, the *actual arrival* and the *arrival gate*, the values given by the arrival airport website have been considered more trustworthy than the ones of the departure airport website.
In particular, the values of the attributes have been chosen as follows:

- *Flight code*: since all the records inside a block had the same value, then there was no need to apply the above mentioned strategies
- *Scheduled departure*, *departure gate*, *scheduled arrival*, *arrival gate*: the *cry with the wolves* and the *trust your friends* strategies have been applied as described above. If the resulting value was missing, then the *null* value has been chosen
- *Actual departure*, *actual arrival*: the *cry with the wolves* and the *trust your friends* strategies have been applied as described above. If the resulting value was missing, then the *take the information* strategy has been used to discover whether the flight had been cancelled or not. For the cancelled flights, the *Cancelled* keyword has been used instead of the *null* value.

Note that the *source* attribute has been used just to identify the trustworthy sources and it has not been included in the final dataset because its records have been generated by fusing the records coming from the various sources.
After the data fusion phase has been completed, the final dataset was nothing but the set of the resulting records from each block, which were expected to be of a higher quality than the original one with respect to the 3 considered data quality dimensions.

5.3 Error analysis

Since the whole data improvement process could have introduced errors, two different error detection methods have been applied.
The first method had the purpose of ensuring that the implementation of all the data improvement phases worked as expected. To accomplish this task, 50 flights randomly selected from the dataset have been manually analyzed.
The second method aimed at evaluating the effectiveness of the data improvement process by comparing the gold standard with the final dataset. This has been carried out by comparing every flight of the gold standard with the corresponding one of the final dataset and every couple of their attributes has been compared. The overall relation error was about 10.8%, where the error has been computed as follows:

$$\frac{\text{number of mismatches}}{\text{number of compared values}}$$

where the number of compared values corresponds to the total number of non-missing values in the gold standard. The results shown in Table 7 demonstrate that almost all the values were correct, except for the ones of the *actual departure* attribute where only 40% of its values were correct. Finally, from Table 8

it is possible to assert that approximately 38.21% of tuples of the final dataset were fully correct, 54.69% had only a wrong attribute (with high probability, it was the *actual departure* attribute), while only about 7.10% had more than a wrong value.

Attribute	**Error**
Flight code	0%
Scheduled departure	0.97%
Actual departure	60.3%
Departure gate	0.04%
Scheduled arrival	1.01%
Actual arrival	7.56%
Arrival gate	0.76%

Table 7: Attribute error compared with the gold standard

Number of tuples	**Number of wrong attributes**
1141	0
1633	1
195	2
15	3
1	5

Table 8: Tuples error compared with the gold standard

6 Final data quality assessment

The dataset obtained at the end of the improving phase has been assessed another time by computing the metrics related to the same data quality dimensions studied in Section 4. Concerning the completeness dimension, Tables 9 and 10 show that the results have significantly improved. In fact, before the data quality improvement phase, only 23.94% of the tuples were complete, while now such percentage grew up to 70.13%. In addition, also the completeness at attribute level increased its value: the one of the *departure gate* attribute has varied from 36.72% to 83.82%, the one of the *arrival gate* attribute from 35.77% to 79.88%, the one of the *scheduled departure* attribute from 66.17% to 99.56%, while the one of the *scheduled arrival* attribute from 66.37% to 99.56%.
Another observation can be done regarding the overall relation completeness, which changed its value from 70% to 93.17%.
As regards the consistency dimension, little improvements have occurred: the C2, C3 and C4 constraints changed their values from 93.47%, 97.97% and 98.42% up to 100%, 99.95% and 99.28%, respectively.
In conclusion, the duplication dimension has remarkably changed from 95.52% to 0%, which means that in the new dataset there are no duplicated flights.

Number of tuples	**Completeness**	**Missing values**
4	14.29%	6
164	42.86%	4
130	57.14%	3
5483	71.43%	2
4618	85.71%	1
24413	100%	0

Table 9: Tuple completeness

Attribute	Completeness
Flight code	100%
Scheduled departure	99.56%
Actual departure	95.81%
Departure gate	83.82%
Scheduled arrival	99.56%
Actual arrival	93.54%
Arrival gate	79.88%

Table 10: Attribute completeness

ID	Constraint	Consistent	Inconsistent
C1	The departure and the arrival airports must be different	100%	0%
C2	The cancelled flights must not have a time for the *actual departure* and the *actual arrival* attributes	100%	0%
C3	*scheduled departure* < *scheduled arrival*	99.95%	0.05%
C4	*actual departure* < *actual arrival*	99.28%	0.72%
C5	*scheduled departure* + min. flight time < *scheduled arrival*	100%	0%
C6	*actual departure* + min. flight time < *actual arrival*	99.99%	0.01%

Table 11: Consistency

References

[1] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, 41(3), July 2009.

[2] Fatimah Sidi, Payam Hassany Shariat Panahy, Lilly Affendey, Marzanah A. Jabar, Hamidah Ibrahim, and Aida Mustapha. Data quality: A survey of data quality dimensions. 08 2013.

[3] Xin Dong and Felix Naumann. Data fusion–resolving data conflicts for integration. pvldb. *PVLDB*, 2:1654–1655, 08 2009.