Reference Sources in Clearing Customer Data: Conclusions from a R&D Project

Mariusz Sienkiewicz[1]

[1]Poznan University of Technology, Poznań, Poland

Abstract

The digitization and virtualization of many aspects of life pose a question for many organizations regarding customer identification. Problem is extremely important in the context of financial institutions (FI), where customer identification is related to a number of aspects of the company's operation and the products and services provided. The problem of unambiguous customer identification consists of data errors, dirty data and duplicate records describing the customer. It is estimated that 1% to approximately 5% of FI data are affected by errors. The scope of data collected by institutions about their clients is enormous and results from many needs. Each of these needs may require a different scope of data and expect different levels of quality. Regardless of the needs for data collection and processing, certain data is particularly important and important - we are talking about data allowing for unambiguous customer identification. In this article, we will pay special attention to the data set that allows for unambiguous customer identification.

Keywords

data cleaning, deduplication, dictionary cleaning, geocoding

1. Introduction

The issue of data cleaning has been raised many times in the literature [1, 2, 3, 4, 5, 6, 7]. There are various suggestions for data error detection and deduplication based on e.g. on the methods of comparing the text, corudsorbing or classification. Based on the experience from a project implemented for a large financial institution, we will present a concept of detecting and correcting customer identification data. This is the first article based on the work carried out on a large (over 2 million records) database of real physical and legal clients.

Due to the nature of their business, financial institutions pay special attention to unambiguous and complete identification of clients. Customer identification is of great importance both for IF collecting data and providing products and services, as well as for customers due to e.g. the security of funds. Financial institutions and financial market regulators pay great attention to the quality of the collected and processed data [8, 9]. Many financial institutions have extensive data management and data quality management systems, and use various techniques to detect errors and correct them in the collected and processed data. These are mechanisms based on specialized software supporting the detection of defects and on customer service procedures focused on the correctness of data, which can be treated as croudsorcing.

Despite the mechanisms used, errors are identified in the data describing clients collected and processed by financial institutions. This is due to: 1) the long history of IT systems, 2) numerous system migrations, 3) acquisitions on the financial market, 4) human errors, 5) intended actions (e.g. attempts to extort financial resources). The effectiveness of the procedures created for the sales force is limited and depends on many factors, and it is not the subject of this article.

Clean and standardized data are required in many areas of data processing in financial institutions, including 1) risk models, 2) security mechanisms, 3) offer and sales support models, 4) ML-based solutions, 5) data deduplication.

The standard data deduplication pipeline [10, 11, 12, 13] assumes that the data delivered to the pipeline is cleared (eg. no zero values, no spelling errors, unified hashes). Unfortunately, this assumption in real projects cannot be guaranteed, especially in the financial sector. There are typos, missing values, inconsistent values in the attributes that store personal data, institution names and addresses. Moreover, not all natural identifiers are reliable. It should also be taken into account that the financial market is largely regulated by law. Interpretation of the current legal regulations and the security practices applied by FI limit the possibility of making changes to customer data and thus the possibility of improving the data. In addition, the client, in accordance with the provisions of contracts concluded by financial institutions, is obliged to ensure that the data made available to the financial institution is up-to-date and correct. Despite the efforts made by financial institutions and the obligations imposed on clients, observations from a project carried out at a large financial institution show that errors in data occur and constitute a significant obstacle for the

Published in the Workshop Proceedings of the EDBT/ICDT 2022 Joint Conference (March 29-April 1, 2022), Edinburgh, UK

mariusz.sienkiewicz@doctorate.put.poznan.pl (M. Sienkiewicz)
URL (M. Sienkiewicz)
0000-0002-1665-4928 (M. Sienkiewicz)

organization.

This article is a continuation of [14] focusing on error detection and improvement of identification data. In particular, we present our experiences and conclusions from the use of reference data sources for error detection and cleaning of customer data (Section 2). We present conclusions regarding cleaning and standardization of address data (Section 3). Final conclusions were drawn in Section 4. Note that this article presents the results of an actual research and development project, and therefore not all details may be disclosed as they are treated as company know-how.

2. Customer identification data

The scope of data collected by financial institutions is wide. A data collection vector describing a single customer can include more than 1000 features. These can be contact, socioeconomic, behavioral data, e.g. product use, transaction data, property ownership status, communication channels used, etc. Data relating to individual features may be dirty. Regardless of the length of the collected data vector, basic identification data are the most important. Of course, the design of IT systems most often ensures the existence of a unique artificial system key that distinguishes records, but from the point of view of a financial institution, it is important to identify all customer instances in order to consolidate knowledge about it.

As a result of the project work, based on the knowledge of the financial institution's experts, a small subset of data was determined, which is particularly important in identifying the client. The basic identification data include: 1) natural key from the population or business registration system, 2) name and surname or name of the entity, 3) document ID, 4) legal form of the entity.

2.1. Detection of identification data errors

Basic identification data errors can be detected using a range of algorithms and tools, i.e. regular expressions, patterns, dictionaries, calculation rules (standard data cleaning mechanisms). However, legal and regulatory constraints significantly limit the use of cleaning mechanisms. Most often, modification is possible after confirming the correctness of these identification data with the customer. Data requirements and availability change over time. In a project implemented for a large financial institution, the researched database of projects has over 2 million records. The entire production customer database is much larger. Verifying the correctness of identification data for the entire customer base is: 1) costly, 2) burdened with image risk, 3) long-term, 4) burdened with human error. Moreover, due to the lack of up-to-date contact details, it may not be possible to reach some of the customers.

As a result of work on deduplication, pairs with a very high degree of similarity are identified for which there are differences in customer identification data. Determining whether the identification data set is correct and which data set is important for the customer may consist of: 1) verification of the data set with the customer - as indicated earlier, this is not an attractive solution from the point of view of the enterprise scale, 2) checking the data on the basis of a reference data sources.

2.2. Verification of identification data

The possibility of mass verification of identification data for a large financial institution seems to be an attractive solution. Access to state registers containing basic data describing the client allows you to verify whether the set of identification data is correct. Verification of identification data with the use of registry data allows to clearly indicate errors in the data and thus to precisely improve the data, which in the case of financial institutions is extremely important.

State registers such as 1) population registration system, 2) register of business activity records should be treated as the reference source of data allowing for the verification of the correctness of the identification data held. Access to individual registers is regulated by law and not all entities can use them equally, and access may be payable. In the project conducted for a large financial institution, 1) the population register and 2) the register of economic activities were used.

2.2.1. Contents of the population register

The population register contains a number of personal data of citizens of a given country. In the case of the Polish register, it is about 30 items. The following are particularly useful for the verification of identification data:

- number ID,
- previous number ID (if changed),
- surname and first names,
- family name,
- the previous surnames and first names with the date of their change and the name of the office that made the change,
- names and surnames of parents (in the case of data change: date and name of the office that made the change),
- date and place of birth (in the case of data change: date and name of the office that made the change),
- country of birth,

- sex (in case of data change: date and name of the office that made the change),
- series and number of the last ID card, its expiry date and the name of the office that issued the ID card,
- series and number of the last passport and its expiry date,
- date of death or the date the body was found, the number of the death certificate and the registry office which drew up the record.

2.2.2. Contents of the register of economic activities

The register of economic activities contains a number of data concerning entities operating in a given country. In the case of the Polish register, it is about 60 items. Particularly useful for the verification of legal entities' identification data are:

- number ID,
- name,
- short name,
- date of creation,
- date of commencement of activities,
- registered office address,
- legal form,
- type of business,
- termination date.

On the basis of the indicated registers, the correctness of the identification data held by the financial institution was tested. In the case of natural persons, these were: 1) the number of the population registration system, 2) first name, 3) surname. For business entities, the following were examined: 1) business registration number, 2) name of the entity, 3) legal form of the business. In the case of legal entities, the access to data is wide and it was possible to verify all entities subject to registration.

With regard to the verification results:

- records were marked where the set of identification data was correct,
- in the case of identified pairs of similar records, where one of the records was confirmed in the reference database, it was possible to decide to create a pair despite differences in identification data, e.g. a different value of one of the compared features,
- designation of a limited set that requires verification in contact with the customer.

The obtained results were verified by experts of the financial institution and proved that the applied cleaning method was adequate to the cleaning problem under consideration. On a representative sample of the records of natural persons from the created pairs, where there was one difference in the identification data, approx. 87% were confirmed to be correct based on the population register.

3. Cleaning of address data

Address data, right after customer identification data, constitute an important element of data in many enterprises, especially in financial institutions (mainly due to numerous information obligations that IFs are obliged to fulfill in a letter form). Designing application interfaces for entering addresses very often, for various reasons, does not have implemented data validation mechanisms. Failure to implement validation rules causes numerous errors in the data. The existence of validation rules does not free the system from the problems related to the purity of address data, as the names of towns and streets may change.

3.1. Address reference data

There are reference databases. These are dictionary systems describing the territorial subordination of a given country. Most often they are organized in the form of hierarchical dictionaries from the largest (province) to the smallest (street) territorial unit. The use of these dictionaries using one of the similarity methods can be used to detect errors in the address data. Due to possible abbreviations, renaming and data errors, the use of territorial dictionaries for validation and improvement of addresses is difficult, especially when we are dealing with a large database of institutions with a long history of functioning and numerous system migrations.

3.2. Geocoder as a tool for standardization of andres data

There are geocoder tools on the market that allow you to efficiently verify the correctness of the address. The most common result of address geocoding is a standardized record of the geocoded address along with the geographic position (longitude, latitude) and quality of match. Geocoders work on the basis of text parsing mechanisms and similarity algorithms - hence the measure of matching, which shows how exactly the geocoded address matches the pattern. Territorial dictionaries are often used as a pattern.

It would seem that since cleared and standardized data is required for the deduplication process, a geocoder type tool is an ideal solution for data cleansing. As mentioned before, the geocoder operates on the basis of text similarity testing methods that the geocoder supplier treats as a trade secret. The project uses a commercial solution

provided by a supplier who has been developing the address base of the territorial area covered by the project for many years and geocoding algorithms, providing solutions for business and individual customers. The tools used by FI constitute a trade secret and cannot be disclosed. In addition, record similarity measures based on text comparison are often used to compare data in the deduplication process. When performing deduplication on geocoded data, you should be aware that the compared data in the previous cleansing and standardization step was established on the basis of some unknown measure of text similarity. Since the geocoder returns some measure of match and is not always able to correctly match the correct address, it is questionable whether comparing records with the measure of similarity of the data text after geocoding is appropriate.

In the implemented project, we decided to use the address data without geocoding them and testing their similarity with the measure of the similarity of the text. Thanks to this approach, the obtained result of comparing records in the deduplication process is not disturbed by the use of indirect rough search processes.

4. Conclusion and Future work

Based on a project implemented for a large financial institution:

- The possibility and usefulness of data from state registers for data correctness verification has been positively verified.
- Access to the data contained in the population register is difficult, and it may turn out to be impossible for entities from outside the financial market or public services.
- Business registers are open but there may be a fee to access them.
- Confirmation of identification data on the basis of state registers is possible for an entity operating on a national scale, international entities would require access to state registers of various countries.
- Not all economic entities (some forms of activity) are included in the business records (they do not require registration),
- There is no confirmation of non-resident data in the population register database.
- The use of territorial dictionaries as a source of reference data requires building mechanisms based on comparing the similarity of the text. Due to the occurrence of abbreviations of names, errors in data, renaming of address names, the usefulness of a solution built solely on the basis of reference data may not be satisfactory.
- The use of a geocoder for cleaning and standardizing address data seems to be justified if in the next steps the data obtained as a result of geocoding will not take part in the comparison of the similarity of the text. The availability of numerical data obtained as a result of geocoding allows for easy comparison of obtained address points based on geographic position. The use of a geocoder in the preparation of data for deduplication requires further research due to the lack of knowledge about the error of geocoding results related to the use of undisclosed methods of comparing texts used in geocoding engines of different suppliers.

Acknowledgements. The work of Mariusz Sienkiewicz is is supported by the Applied Doctorate Scholarship no. DWD/4/24/2020 from the Ministry of Education and Science and additionally the project is supported by a grant from the National Center for Research and Development no. POIR.01.01.01-00-0287/19.

References

[1] X. Chu, Data cleaning, in: S. Sakr, A. Y. Zomaya (Eds.), Encyclopedia of Big Data Technologies, Springer, 2019. URL: https://doi.org/10.1007/978-3-319-63962-8_3-1. doi:10.1007/978-3-319-63962-8_3-1.

[2] E. K. Rezig, Data cleaning in the era of data science: Challenges and opportunities, in: 11th Conference on Innovative Data Systems Research, CIDR 2021, Virtual Event, January 11-15, 2021, Online Proceedings, www.cidrdb.org, 2021. URL: http://cidrdb.org/cidr2021/papers/cidr2021_abstract09.pdf.

[3] X. Chu, I. F. Ilyas, S. Krishnan, J. Wang, Data cleaning: Overview and emerging challenges, in: F. Özcan, G. Koutrika, S. Madden (Eds.), Proceedings of the 2016 International Conference on Management of Data, SIGMOD Conference 2016, San Francisco, CA, USA, June 26 - July 01, 2016, ACM, 2016, pp. 2201–2206. URL: https://doi.org/10.1145/2882903.2912574. doi:10.1145/2882903.2912574.

[4] G. Y. Lee, L. Alzamil, B. Doskenov, A. Termehchy, A survey on data cleaning methods for improved machine learning model performance, CoRR abs/2109.07127 (2021). URL: https://arxiv.org/abs/2109.07127. arXiv:2109.07127.

[5] E. Rahm, H. H. Do, Data cleaning: Problems and current approaches, IEEE Data Eng. Bull. 23 (2000) 3–13. URL: http://sites.computer.org/debull/A00DEC-CD.pdf.

[6] O. Azeroual, Data wrangling in database systems: Purging of dirty data, Data 5 (2020) 50. URL: https://doi.org/10.3390/data5020050. doi:10.3390/data5020050.

[7] M. A. Hernández, S. J. Stolfo, Real-world data is dirty: Data cleansing and the merge/purge problem, Data Min. Knowl. Discov. 2 (1998) 9–37. URL: https://doi.org/10.1023/A:1009761603038. doi:10.1023/A:1009761603038.

[8] T. P. F. S. Authority, Recommendation d. concerning the management of information technology areas and security of the ict environment in banks, https://www.knf.gov.pl/knf/pl/komponenty/img/Rekomendacja_D_8_01_13_uchwala_7_33016.pdf, 2013.

[9] O. J. of the European Union, Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), https://eur-lex.europa.eu/eli/reg/2016/679/oj, 2016.

[10] A. Colyer, The morning paper on An overview of end-to-end entity resolution for big data, https://blog.acolyer.org/2020/12/14/entity-resolution/, 2020.

[11] A. Simitsis, P. Vassiliadis, T. K. Sellis, State-space optimization of ETL workflows, IEEE Transactions on Knowledge and Data Engineering 17 (2005) 1404–1419.

[12] G. Papadakis, D. Skoutas, E. Thanos, T. Palpanas, Blocking and filtering techniques for entity resolution: A survey, ACM Comput. Surv. 53 (2020) 31:1–31:42.

[13] G. Papadakis, L. Tsekouras, E. Thanos, G. Giannakopoulos, T. Palpanas, M. Koubarakis, Domain- and structure-agnostic end-to-end entity resolution with jedai, SIGMOD Record 48 (2019) 30–36.

[14] M. Sienkiewicz, R. Wrembel, Managing data in a big financial institution: Conclusions from a r&d project, in: C. Costa, E. Pitoura (Eds.), Proceedings of the Workshops of the EDBT/ICDT 2021 Joint Conference, Nicosia, Cyprus, March 23, 2021, volume 2841 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2841/DARLI-AP_9.pdf.