R.2.3 - Data Reconciliation: Approaches and Problems

Team progetto DoUtDes

University of Cagliari, Italy
doutdes.unica@gmail.com

Abstract. Data Reconciliation (DR) is a process aimed to match records that are stored on different databases, for instance in order to correlate different activities of the same user (ID reconciliation). Such a process operates by joining non-key dataset fields by using matching algorithms able to face problems related to the differences between the involved datasets, in terms both of structure and potential presence of data errors. The approximate matching algorithms are configured (parameter settings) according to the data scenario, and its computational complexity depends on the needed number of comparisons between datasets. For this reason, such approaches can not be taken into account in data scenarios that involve very large datasets, due to the unsustainable computational cost. The aforementioned technical problems are flanked by the increasingly rigorous rules for the privacy protection, which in many cases do not make applicable these approaches. A case in point of this is the European General Data Protection Regulation (GDPR), in force from May 2018, with the purpose to protect the EU citizens' personal data, considering that in almost all cases it does not allow the commercial operators to cross-reference their users' data in order to profile them in greater detail.

Keywords: Data Reconciliation · Business Intelligence · Decision Support System · Privacy · Algorithms

Acknowledgements. This research is partially funded and supported by the Aut. Reg. of Sardinia cluster project *"DoUtDes. Trasferimento di tecnologie e competenze di Business Intelligence alle aziende dei settori innovativi e tradizionali"*,
CUP: F21B17000850005 (POR-FESR SARDEGNA 2014-2020).

1 Introduction

One of the main objectives of a trader is to define as detailed a profile as possible of their customers, based on their behavior over time [1] (e.g., products and services purchased by them). The reason for this is given by the fact that detailed customer

profiles allow them to deploy very targeted and effective marketing choices, with the consequent economic advantage [2].

The level of detail in user profiling can be significantly increased if more commercial operators cross the user information stored in their databases, carrying out a DR operation [3]. In addition to the advertising and marketing perspective , such operations must be considered under a regulatory perspective, taking into account the nature of information exchanged across different datasets. The recent European General Data Protection Regulation (GDPR) introduces a series of restrictions with respect to the past, making some previously permitted practices illegal [4].

It should be noted how the available information about the users are in constant and exponential growth, since their number is directly related to the increase in E-commerce activities, which are further increased from some time due to the COVID-19 world emergency [5, 6].

This scenario has been investigated by several studies, which have confirmed the exponential growing in the E-commerce activities, as shown in Figure 1 that reports a study performed by eMarketer [1]. It shows how the E-commerce businesses should get a 265% growth rate, from USD 1.3 trillion in 2014 to USD 4.9 trillion in 2021.

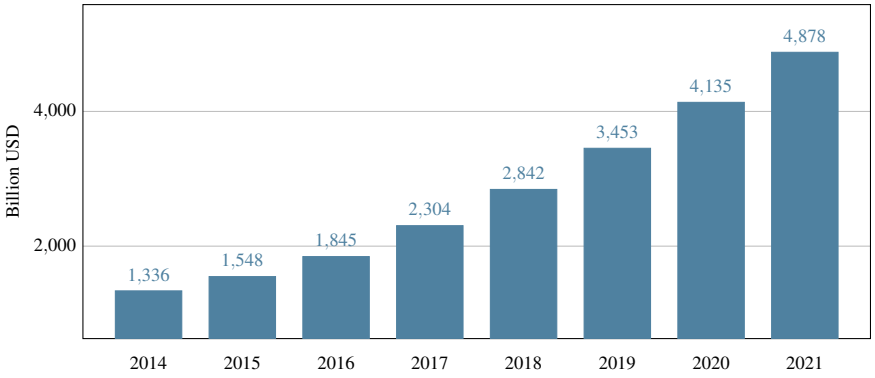

Fig. 1. Total Retail Sales Worldwide 2014 ÷ 2021

This information-rich scenario has further evolved into the so-called Social Commerce, where networking websites such as Facebook, Instagram, and Twitter have been exploited in order to promote and sell goods and services [7]. Summarizing what has been said so far, we are faced with a scenario unthinkable until a few years ago, a very huge number of information about people that can be crossed, becoming powerful sales tools, but it is necessary to filter these activities on the basis of the laws in force, a problem to deal since the origins of E-commerce [8].

According to the literature, it should be added that the DR techniques are also used to verify the data correctness after their migration from a database to another ones, a

[1] https://www.emarketer.com

similar comparison process between data sources, but performed with a different objective [9].

2 Literature Review

Nowadays, anyone who operates using network services, how it happens, for instance, in the E-commerce environment, is associated with an identifier (ID), which depending on the context can be explicit (personal data) or implicit (network address) [10]. The literature indicates a great effort in this research direction, proposing solutions based on different techniques and strategies (e.g., machine learning, artificial intelligence, etc.) aimed to define effective and unique IDs on the basis of one or more data sources. According to the needs, such IDs can be referred to users, products, brands, and so on [11].

By way of example, in a real-world common context, the available sources of data are used to generate as accurate as possible customer profiles. This operation requires also the joining on secondary database fields, which are used to relate information from the same customer, when primary keys are not available [3].

When very huge databases are involved in the data reconciliation process, a canonical pairwise comparison between records is not feasible, due to the high computational cost that this operation requires. In order to tackle this problem, the literature has made available for a long time different methods and strategies capable of reducing the computational load.

A case in point is the work in [12], where the authors propose a modified merge-sort procedure aimed to reduce the computational load in the context of large data through a duplicate elimination technique. In another work in [13], the task related to the correlation of information from different databases have been taken into consideration in order to identify distinct users that are present in different data sources, and in such a context the authors propose a computational load reduction method.

The techniques mentioned above represent only some of the best known, but also many other recent techniques/strategies related to the databases (then not directly related to the DR) can be profitably used for this purpose. For instance, in the work in [14], the authors propose the implementation of distributed join algorithms in the context of systems with several thousand cores, which are connected through a low-latency network. In the work in [15], an Artificial Bee Colony Algorithm based on Genetic Operators has been proposed by the authors with the aim to optimize the queries between databases, reducing the computational effort.

3 Legal Aspects

Regardless of what concerns the DR approaches, they must be implemented in compliance with the laws in force, primarily the aforementioned GDPR [4]. By way of example, the GDPR does not specify, explicitly, how long can be the retention periods for personal data (e.g., how long a commercial operator can store the personal data related to the customers), but it states that such data may only be ”*kept in a form which*

permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed".

Other GDPR articles such as "*processed lawfully, fairly and in a transparent manner in relation to the data subject*" or "*collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes*".

It should be noted how the profiling represents a special case of automated processing of personal data in the *Article 4* of the GDPR, which describes it as "*any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces legal effects concerning him or her or similarly significantly affects him or her*". It goes on to say "*In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision*".

Such rules and exceptions can raise doubts about what is allowed or not, as the involved variables are strictly dependent on the scenario in which one operates.

4 Conclusions

In light of the foregoing information, we can observe how the techniques related to the Data Reconciliation are used for different purposes ranging from the verification of the correctness of information after a database migration, to the identification of distinct users/customers, on the basis of information stored on different databases (ID Reconciliation). At the same time we have seen how the technical possibilities must take into account the laws in force.

For this reason, we can conclude by saying that it is not possible defining an unique guideline, with regard to the DR techniques applications, as it depends on the context in which they are used. This also in consideration of the value that the concept of privacy has assumed in recent years, as proven by many studies that indicate how the new GDPR rules are having a huge positive impact on consumer opinion, with regard to personal data being collected and stored by public and private entities.

References

1. Eke, C.I., Norman, A.A., Shuib, L., Nweke, H.F.: A survey of user profiling: State-of-the-art, challenges, and solutions. IEEE Access **7** (2019) 144907–144924
2. Yang, W.S., Dia, J.B., Cheng, H.C., Lin, H.T.: Mining social networks for targeted advertising. In: Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06). Volume 6., IEEE (2006) 137a–137a
3. Cochinwala, M., Kurien, V., Lalk, G., Shasha, D.: Efficient data reconciliation. Information Sciences **137**(1-4) (2001) 1–15
4. Goddard, M.: The eu general data protection regulation (gdpr): European regulation that has a global impact. International Journal of Market Research **59**(6) (2017) 703–705

5. Bhatti, A., Akram, H., Basit, H.M., Khan, A.U., Raza, S.M., Naqvi, M.B.: E-commerce trends during covid-19 pandemic. International Journal of Future Generation Communication and Networking **13**(2) (2020) 1449–1452
6. Barnes, S.J.: Information management research and practice in the post-covid-19 world. International Journal of Information Management **55** (2020) 102175
7. Bugshan, H., Attar, R.W.: Social commerce information sharing and their impact on consumers. Technological Forecasting and Social Change **153** (2020) 119875
8. Ackerman, M.S., Cranor, L.F., Reagle, J.: Privacy in e-commerce: examining user scenarios and privacy preferences. In: Proceedings of the 1st ACM conference on Electronic commerce. (1999) 1–8
9. Bakhtouchi, A.: Data reconciliation and fusion methods: A survey. Applied Computing and Informatics (2019)
10. Sharma, S., Rana, V.: Web user identification: a review of approaches and issues. Int. J. Comput. Eng. Technol.(IJCET) **8**(4) (2017) 12–18
11. Zhao, K., Li, Y., Shuai, Z., Yang, C.: Learning and transferring ids representation in e-commerce. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. (2018) 1031–1039
12. Bitton, D., DeWitt, D.J.: Duplicate record elimination in large data files. ACM Transactions on database systems (TODS) **8**(2) (1983) 255–265
13. Hernández, M.A., Stolfo, S.J.: The merge/purge problem for large databases. ACM Sigmod Record **24**(2) (1995) 127–138
14. Barthels, C., Müller, I., Schneider, T., Alonso, G., Hoefler, T.: Distributed join algorithms on thousands of cores. Proceedings of the VLDB Endowment **10**(5) (2017) 517–528
15. Panahi, V., Navimipour, N.J.: Join query optimization in the distributed database system using an artificial bee colony algorithm and genetic operators. Concurrency and Computation: Practice and Experience **31**(17) (2019) e5218