Applying Domain Knowledge for Data Quality Assessment in Dermatology

Nemanja Igić[1(✉)], Branko Terzić[1], Milan Matić[2], Vladimir Ivančević[1], and Ivan Luković[1]

[1] Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
{nemanjaigic,branko.terzic,dragoman,ivan}@uns.ac.rs
[2] Faculty of Medicine, University of Novi Sad, Novi Sad, Serbia
milan.matic@mf.uns.ac.rs

Abstract. The Dermatology Clinic at the Clinical Center of Vojvodina, Novi Sad, Serbia, has actively collected data regarding patients' treatment, health insurance and examinations. These data were stored in documents in the comma-separated values (CSV) format. Since many fields in these documents were presented as free form text or allow null values, there are many data records that are inconsistent with the real-world system. Currently, there is a large need for an analytic system that can analyze these data and find relevant patterns. Since such an analytic system would require clean and accurate data, there is a need to assess data quality. Therefore, a data quality system should be designed and built with a goal of identifying inaccurate records so that they can be aligned with the real-world state. In our approach to data quality assessment, the domain knowledge about data is used to define rules which are then used to evaluate the quality of the data. In this paper, we present the architecture of a data quality system that is used to define and apply these rules. The rules are first defined by a domain expert and then applied to data in order to determine the number of records that do not match the defined rules and identify the exact anomalies in the given records. Also, we present a case study in which we applied this data quality system to the data collected by the Dermatology Clinic.

Keywords: Dermatology · Data quality assessment · Domain knowledge application

1 Introduction

Collection and analysis of multi-year data are important activities in many large systems. Finding relevant patterns, analyzing trends and building good predictive models are the key elements of good decision making. In medicine, analytic systems are very important in identifying key indications in disease diagnosis [1] and the right prevention is essential both financially and in terms of patient health. There are many examples of advanced analysis of medical data and extraction of meaningful information, such as using neural networks in diagnosis [2], using big data analysis to find business value in health care [3], and using data mining in the treatment of heart diseases [4].

I. Czarnowski et al. (eds.), *Intelligent Decision Technologies 2017*,
Smart Innovation, Systems and Technologies 73, DOI 10.1007/978-3-319-59424-8_14

To get maximum benefits from using an analytic system, the analyzed data must be correct and trustful. Therefore, each anomaly in data records needs to be detected and fixed before any analysis is performed. There must be a well-defined process that aims to evaluate data and measure their quality by using the provided metrics. Failing to define such a process may result in poor and inaccurate results, which can lead to making bad and, in the medical domain, potentially dangerous conclusions.

Data quality assessment generally includes evaluation of data quality based on some predefined metrics. By using a generic model of data quality assessment that relies on well-defined objective and subjective metrics based on questionnaires [5], we can perform unbiased data quality assessment. However, in an analysis of data gathered from the Dermatology Clinic at the Clinical Center of Vojvodina (CCV), Novi Sad, Serbia, a more specific model of data quality assessment is required. For this purpose, we focus more on the domain experts' view of data because the domain experts will be interpreting the results of the analytic system that will be built for these data. Thus, this kind of data quality assessment gives emphasis to the domain experts' interpretation of data rather than a general one.

Our main goal is to construct a data quality system that will allow domain experts to easily present their knowledge about the data by offering them a wide range of easily understandable concepts for defining rules. After applying the defined rules to the provided data records, our system will generate the results of data quality assessment in the form of a report. Domain experts can then use the results from the report to conclude which data records are inconsistent with the defined rules. After that, domain experts should be able to investigate the causes of the discovered inconsistencies and suggest the preferable data values.

The system presented in this paper relies on the domain knowledge about the data. Since the Dermatology Clinic is the holder of data, we presume that the experts at the clinic may offer the most relevant interpretation of data. By letting these experts present their view about data as a set of rules, we can easily analyze how much inconsistency there is and which concrete rules trigger negative results for the analyzed data records. In this way, we can determine exact differences between the real world and the data gathered by the Dermatology Clinic. Our hope is that we can get better results by using domain specific metrics that are rule based rather than the generic metrics, which are predefined for any case study, as presented in [5].

Besides Introduction and Conclusion, this paper has four more sections. In Sect. 2, we describe various approaches to data quality assessment, with emphasis on data quality assessment in medicine. In Sect. 3, we describe data gathered from the Dermatology Clinic at CCV. In Sect. 4, we present the architecture of the system that we used to assess data quality. In Sect. 5, we present a case study in which we applied our data quality system to the data gathered by the Dermatology Clinic. In this case study, domain experts defined a set of rules, which were applied to the provided data. Also, we present results of the report generated by our data quality system and the domain experts' conclusions.

2 Related Work

In this section, we first describe various approaches in data quality assessment. Afterwards, we present examples of data quality assessment in medicine.

2.1 Data Quality Assessment Approaches

In [6], there is a comparison between many different data quality methodologies that can be applied in various types of information systems. All those methodologies have a goal to define a set of objective metrics and activities in data quality assessment for the given type of information system. For example, in [7], a system for managing data quality for cooperative information systems is presented. In [8], an example of a modeling data quality process for multi-input and multi-output information system is presented. There are cases where data quality is treated as product quality and product quality procedures are applied [9]. There are also methodologies that are based on questionnaires to gather data on the status of the organizational information quality [10]. In [11, 12], there are procedures on how to manage data quality in big data and large multi-organization systems, which are very popular nowadays. As discussed in [13], there are also approaches where organizational structure is taken into account. All these approaches are using predefined metrics which are easy to apply and interpret, but give more generic results. In our approach, we emphasize domain experts' knowledge in order to get in-depth results of data quality assessment.

Besides defining metrics and processes for data quality assessment, there are approaches in which data mining techniques are used to accomplish the same task. As presented in [14, 15], association rules can be used to find data anomalies based on very high or very low values of confidence. This rule-based approach gives some hidden insights about the data that might not be obvious, but there is also possibility not to emphasize other very important conclusions about the data. On the other hand, our approach gives a wider set of concepts to define different rules, rather than just using association rules. Also, we are more focused on important aspects of data quality assessment from the domain experts' viewpoint, rather than finding hidden knowledge in data.

During our research, within the available literature, we did not find any approaches that are based directly on the domain knowledge from persons who will interpret data after the data are cleaned and transformed.

2.2 Data Quality in Medicine

There are also research studies concerning data quality assessment in the field of medicine. In [16], there is a description of a framework that enables clinical institutions to plan and control the data quality assessment process based on dimensions that were predefined for the clinical domain. In [17], there is a discussion of problems in data quality assessment that are related to integration of various sources in a single warehouse when different naming conventions are used, e.g., for gender and race. The proposed solutions are the standardization of the values and strict database constraints. In [18], a complete process of data quality assessment is proposed with defined metrics that were

proven in practice. Even though these approaches are used in medicine, they are more focused on defining various metrics for this field rather than applying domain knowledge in data quality assessment.

3 Data Sources

In this section, we describe the data collected at the Dermatology Clinic at CCV. The data were gathered during the period from 2010 to 2015. Those data contain information about patients treated at the Dermatology Clinic and they were pulled from the CCV information system. The data are split into several comma-separated values (CSV) documents, based on database tables. More details about these tables are presented in the following subsections.

3.1 Outpatients and Hospital Treated Patients

There are two documents describing the patient's relationship with the clinic. The first document contains data about patients treated at the outpatient department of the clinic, i.e., patients who are not treated at the hospital and are present at the clinic only for the examination. The second document contains data about inpatient treatment. The document that contains data about the outpatient treated patients has 60516 data records and the document that contains data about hospital treated patients has 4507 data records. Both documents have the same data structure. There are 51 attributes, including the examination time, patient's health insurance, basic patient information and information about the doctor performing the examination. Based on the fields from these two documents, the domain experts have defined a set of rules that we used in the data quality assessment case study, described in Sect. 5.

3.2 Examination Reports

There are 14960 examination reports for the given period of time. Each examination report contains data about the patient, doctor, and examination time, as well as a text report, which is created as an unstructured text description. These documents were not used in the presented case study, due to simplicity and ease of its interpretation.

3.3 Medicines and Supplies

This document contains data about medicines and supplies that were prescribed to patients or used during treatment. There are 344676 data records. The attributes from this document describe the medicine or supply, date, patient and quantity. This document was not used in the presented case study, for the same reason as the previous one.

4 Architecture

As presented in Fig. 1, the architecture of our system comprises six components: a rule builder, a tokenizer, a parser, a rule assembler, a data quality assessment (DQA) engine, and a report engine. These components constitute a pipeline and run in the sequential order depicted in Fig. 1. They were implemented in Java using the Spring framework [19] and the Java Data Mining Package [20]. In the remainder of this section, we describe each component, as well as input files and produced report files.

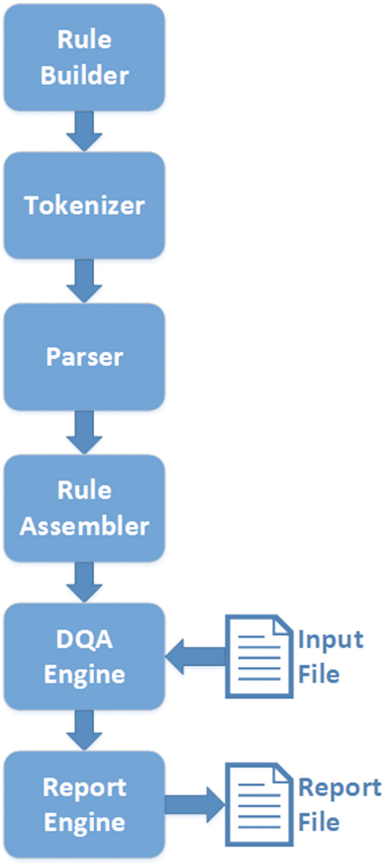

Fig. 1. A presentation of the system architecture

Rule Builder. A domain expert may use the rule builder through its user interface to define a set of rule specifications for the given business rule. In a guided process, the domain expert may use one of the provided rule patterns, select attributes of interest and create rule specification as a relationship between the attributes based on the given

patterns. Rule patterns, which are defined separately by the developer, represent meta-rules through which concrete rules may be defined. Based on the created set of rule specifications, a domain expert may conduct data quality assessment.

Tokenizer. The tokenizer is responsible for extracting tokens from the rule specifications created by the domain expert. Tokens are used to determine to which rule pattern the provided rule specification belongs. Tokens represent predefined labels that describe one concrete concept and are used as a part of rule patterns. Tokens are defined as part of a data quality grammar and each token corresponds to one regular expression [21]. Tokens are recognized in the rule specifications by the corresponding regular expressions. Each rule specification corresponds to a list of tokens. Regular expressions must be ordered from specific to more general since regular expressions are applied to rule specifications sequentially. A part of the rule that satisfies the given rule expression is marked as a token. After that, regular expressions are again applied to the rest of the rule. This procedure continues until the whole rule is tokenized.

Parser. The parser takes a list of tokens for each rule specification as input and checks which rule pattern corresponds to the given list of tokens. Each list of tokens corresponds to one rule pattern. Rule patterns are defined using EBNF notation [22].

Rule Assembler. This component gets, as input from the Parser component, a map where the key is a rule specification and the value is a rule pattern to which the rule specification corresponds. Based on a rule pattern, the rule assembler transforms the concrete rule specification to a Boolean expression, which is later applied to each data record to check if the rule is true for the given record.

Input File. The input file represents a CSV document based on which the data quality assessment is made.

DQA Engine. The Data Quality Assessment (DQA) engine is used to check if the Boolean expressions provided by the rule assembler are true for the given data records. Each Boolean expression is applied to the given data record and, for each pair consisting of a rule specification a data record, true or false is returned as a result, i.e., if the rule specification for the given record is satisfied or not.

Report Engine. For each rule specification - data record pair for which the DQA engine returns false, a template message is generated based on the rule specification and the data record values. Each template message corresponds to the rule pattern that satisfies the given rule specification.

Report File. Compiled template messages from the report engine are appended to the report file, which is presented to the user as the final result. The report file also presents statistics about the data quality assessment process, such as the number of data records that yielded false for the given rule.

5 Case Study

In this case study, we present an example of using the proposed data quality system. We show how to define a rule specification based on the given set of rule patterns. This case study is based on a subset of real business rules that are defined by domain experts from the Dermatology Clinic in CCV and are applied on the data set from that clinic. The goal of this case study is to determine how many, and which, data records are not in accordance with the provided rules.

To define rule specifications, based on which the data quality assessment will be executed, rule patterns need to be defined. Rule patterns are used to define in what way a concrete rule will be specified and to define the rule semantics. Table 1 presents supported rule patterns for this case study. Rule patterns should be defined in the EBNF notation and there are no other limitations regarding their definition. Rule pattern are defined, as presented in Table 1, in the parser component and the pattern semantics, i.e., how the rules defined by the pattern will be presented as Boolean expressions by which data records will be evaluated, are implemented in the DQA engine.

Table 1. Supported rule patterns in the case study

Rule pattern	Description
IS_DATE : = DATE	Checks if the token is date
IS_NULL : = FIELD NOT NULL	Checks if the field is not null
REL_EXP : = NUM_FIELD REL_EXP NUM_FIELD REL_EXPS : = REL_EXP | REL_EXPS AND REL_EXP | REL_EXPS OR REL_EXP	Checks if the relational expression is true. This rule can be applied recursively by using the operators AND and OR
FUN_REL : = FIELDS => FIELDS | FIELDS : = FIELD FIELDS FIELD	Checks if the left hand side attributes are functionally related with the right hand side attributes

To present concrete business rules, the domain expert needs to write rule specifications using the given rule patterns. For example, we can define that medical insurance number is mandatory and write a rule specification as: INS_NUM NOT NULL, where INS_NUM represents the name of the insurance number column. We can also define a rule specification stating that the number of days spent in hospital for outpatients is one: STATUS = A AND NDAYS = 1, where STATUS represents a value defining if the patient has the outpatient status and NDAYS represents the number of days the patient has been hospitalized. We can assume that the information about the patient's clinic substation functionally determines the information about the patient's home town since patients have to be registered at the clinic substation nearest to their place of living. Therefore, we can get a rule specification which is presented as SUBSTATION => HOME_TOWN. After we define the set of rule specifications, we can apply them to our current data set. After running rules over our documents regarding outpatients' data, we get the results presented in Tables 2 and 3.

Table 2. The resulting statistics of applying the given rules over the data set.

Rule	Number of anomalies	Percentage of records with anomaly
INS_NUM NOT NULL	674	1.13%
STATUS = A AND NDAYS = 1	17	0.028%
SUBSTATION => HOME_TOWN	121	0.12%

Table 3. A part of the resulting report file

Row	Rule	Explanation
4	STATUS = A AND NDAYS = 1	When STATUS is A, NDAYS is 1
15	INS_NUM NOT NULL	INS_NUM must not have a null value
28	STATUS = A AND NDAYS = 1	When STATUS is A, NDAYS is 1
455	INS_NUM NOT NULL	INS_NUM must not have a null value
460	INS_NUM NOT NULL	INS_NUM must not have a null value
1499	SUBSTATION => HOME_TOWN	HOME_TOWN must not have different values for the same value of the field SUBSTATION

In Table 2, for each rule specification that has been defined, we present both the number of anomalies and the percentage of records with the corresponding anomaly. For the first rule specification in Table 2, there are 674 anomalous data records in which the insurance number is not specified. This may be the case of an emergency situation when only basic information is entered while the rest of the record is never updated afterwards. There are only 17 cases that do not satisfy the second rule specification in Table 2 and this might be due to a typing error. The third rule specification in Table 2 is not matched within 121 data records. This might happen when patients change their place of living but they do not change their substation yet. Table 1. Supported rule patterns in the case study.

Based on the report segment presented in Table 3, it is possible to see the exact row number where an anomaly occurred and the rule specification that was not satisfied. For example, one rule specification was not satisfied for row 4 since the number of days an outpatient stays in hospital should be 1 but it was 4. In this case, the number of days needs to be set to 1 or the patient's status needs to be set to hospitalized.

6 Conclusion

Using domain knowledge represented as a set of rules can be viewed as a flexible approach to data quality assessment in the domain of dermatology. One of the important advantages of this approach is the absence of fixed and generic metrics in the process of data quality assessment. Also, this approach ensures that the domain expert who will interpret data can make a set of rules through which all anomalies may be found, even if these rules are simple not null rules or complex, domain specific rules. One might

argue that using generic metrics might be easier to implement and use in practice. In certain situations, it may be complex to define rule patterns that can appear, tokens for the rule pattern corpus, and concrete rules. Analyzing system complexity and making a pilot solution is necessary before fully applying this approach. The presented system yielded satisfactory results in our case study about the quality of dermatologic data since we did not have many problems in detecting the needed rule patterns and tokens. The domain expert has easily detected which concrete rules should be applied, but this might not always be the case. This approach also worked very well in our scenario since we needed data interpreted from the domain expert's point of view, so we can build a more accurate and reliable analytic system based on that data. Our future research in this domain will be focused on implementing an autocorrection mechanism of the data records that are marked as false for the given rule, where applicable. Also, we plan to work on a mechanism which will resolve conflicts between rules, if they occur.

Acknowledgements. The research presented in this paper was supported by the Ministry of Education, Science, and Technological Development of the Republic of Serbia under Grant III-44010. The authors are most grateful to Clinical Center of Vojvodina for the provided data set and valuable support throughout the study.

References

1. Kwetishe, D., Osofisan, A.O.: Evaluation of predictive data mining algorithms in Erythemato-Squamous disease diagnosis. IJCSI Int. J. Comput. Sci. Issues **11**(6), 85–94 (2014)
2. Brause, R.W.: Medical analysis and diagnosis by neural networks. Med. Data Anal. **2199**, 1–13 (2001)
3. Ji, Z.: Applications analysis of big data analysis in the medical industry. Int. J. Database Theor. Appl. **8**(4), 107–116 (2015)
4. Shouman M., Turner T., Stocke R.: Using data mining techniques in heart disease diagnosis and treatment. In: Proceedings of the 2012 Japan-Egypt Conference on Electronics, Communications and Computers, pp. 173–177 (2012)
5. Pipino, L.L., Lee, Y.W., Wang, R.Y.: Data quality assessment. Commun. ACM Support. Commun. Build. Soc. Capital **45**(4), 211–218 (2002)
6. Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: Methodologies for data quality assessment and improvement. ACM Comput. Surv. **41**(3), 16 (2009). Article No. 16
7. Scannapieco, M., Virgillito, A., Marchetti, C., Mecella, M., Baldoni, R.: The DaQuinCIS architecture: a platform for exchanging and improving data quality in cooperative information systems. Inf. Syst. **29**(7), 551–582 (2004)
8. Ballou, D.P., Pazer, H.L.: Modeling data and process quality in multi-input, multi-output information systems. Manage. Sci. **31**(2), 150–162 (1985)
9. Ballou, D.P., Wang, R.Y., Pazer, H., Tayi, G.K.: Modeling information manufacturing systems to determine information product quality. Manage. Sci. **44**(4), 462–484 (1998)
10. Lee, Y.W., Strong, D.M., Kahn, B.K., Wang, Y.W.: AIMQ: a methodology for information quality assessment. J. Inf. Manage. **40**(2), 133–146 (2002)
11. Laudon, K.C.: Data quality and due process in large interorganizational record systems. Commun. ACM **29**(1), 4–11 (1986)
12. Cai, L., Zhu, Y.: The challenges of data quality and data quality assessment in the big data era. Data Sci. J. **14**, 2 (2015)

13. Silvola, R., Harkonen, J., Vilppola, O., Kropsu-Vehkapera, H., Haapasalo, H.: Data quality assessment and improvement. Int. J. Bus. Inf. Syst. **22**(1), 62–81 (2016)
14. Hipp, J., Guntzer, U., Grimmer, U.: Data quality mining. In: Proceedings of the 6th ACM Sigmod Workshop on Research Issues in Data Mining and Knowledge Discovery (2001)
15. Farzi, S., Baraani, D.A.: Data quality measurement using data mining. Int. J. Comput. Theor. Eng. **2**(1), 1793–8201 (2010)
16. Nahm, M.: Data quality in clinical research. In: Richesson, R.L., Andrews, J.E. (eds.) Clinical Research Informatics, pp. 175–201. Springer, London (2012)
17. Bae, C.J., Griffith, S., Fan, Y., Dunphy, C., Thompson, N., Urchek, J., Parchman, A., Katzan, I.L.: Challenges of data quality in medical informatics data warehouses. EGEMS (Wash DC) **3**(1), 1125 (2015)
18. Zozus M.N., Ed Hammond, W., Green, B.B., Kahn, M.G., Richesson, R.L., Rusincovitch, R.A., Simon, G.E., Smerek, M.M.: Assessing data quality for healthcare systems data used in clinical research. NIH Health Care Systems Research Collaboratory
19. Spring – Spring Framework. http://spring.io/
20. JDMP – Java Data Mining Package. http://jdmp.org/
21. Karttunen, L., Chanod, J.P., Grefenstette, G., Schiller, A.: Regular expressions for language engineering. Nat. Lang. Eng. 1–24 (1997)
22. Scowen, R.S.: Extended BNF — a generic base standard. In: Proceedings of the Software Engineering Standards Symposium (1993)