Data Cleansing According to Ontology-Based Axioms

Morteza Mahdiani
Amirkabir University of Technology(Tehran Polytechnic)
m.mahdiani@aut.ac.ir

Abstract

We need to clean and cleans data before data processing. According to our problem and data set, data cleaning contains some steps to obtain correct data while we need data cleansing to gain validated and verified records which mostly relates to semantic analysis. There are various methods for preprocessing datasets. Among these methods, we used a beneficial approach that applies ontology-based concept as a generic way to clean and cleans data. By introducing an ontology-based data cleaning and cleansing method, we could semantically preprocess the data by means of axioms that were defined in our ontology-based. This kind of preprocessing led us to better results in the processing phase where we made a semantically unclean data from actual data and made them clean and cleans with our approach. After that, we processed them with a similar method and the outputs demonstrated that our approach works well.

Introduction

It has been some decades since we have started to feel an increasing need to handle different kinds of data in scientific problems. Most scientific research struggle with processing related data, on the other hand, we are facing a question that is: are we using validated and verified data for our problem?

First, our data should be validated. Second, it should be verified. Verification and validation (V&V) are the activities performed during a software development project to ensure that the right system is developed and it meets the expectations of its customers (validation) and that this developed system is correct and conforms to its specifications (verification) [1].

There are a number of ways to achieve validated and verified data equal to cleaning and cleansing procedures in the preprocessing phase of data analysis. In this paper we refer to cleaning as steps that are needed to have a well-structured data, including filling missing values, finding duplicates etc. which are mostly syntactic errors. Then we refer to cleansing as finding meaningful records, eliminating records that don’t have a definition in the specific domain etc. which are mostly semantic errors.

Our approach

While most of the represented methods belong to the cleaning phase, we used an ontology-based concept to achieve cleansing goals that mostly relate to semantic analysis. We used COGNIBASE[4] as an ontology[15] to develop our ontology-based that consists of five essential elements. The most important element for us is an axiom class that directly helps us in detecting and correcting the incorrect and irrelevant semantic records in the datasets that are being manipulated.

Finally, for evaluating our approach, we used an SVM classification algorithm[2]. We started with summer Olympic medalist from 1896 to 2008 dataset[11]. The labels include bronze, silver, and gold that would create a multivariate classification problem for us. We made dirty data[1] from this dataset with the described algorithm that needs to be cleansed. After that, we applied our approach to have cleansed data. In the end, three models were obtained from these three datasets (actual, dirty and cleansed). After comparing the predicted values of the same test vector from these models we found a significant difference between predicted values of the model that was derived from actual data and the model that was derived from dirty data. Nevertheless, the results of the model derived from actual data and those of the model from cleansed data don't show significant difference.

Previous work and our approach

We want to find an approach for data cleansing that finds irrelevant and incorrect records that mostly relate to semantic analysis in the preprocessing phase. To overcome this problem we used an ontology-base which provided us with a set of concepts, taxonomic and non-taxonomic relations and axioms in a specific domain[4]. A number of previous work have used the idea of applying ontology in data cleaning. Among them, an ontology-based approach for data cleaning[5] uses ontology for mapping operations of data cleaning between different datasets. Xin Wang, Howard John Hamilton and Yashu Bither introduced a framework for detecting and correcting errors in databases[8] that uses an ontology-base. Another work[9] focused on differences in terminologies and finding a solution for detecting and solving them based on linguistic knowledge provided by a domain ontology.

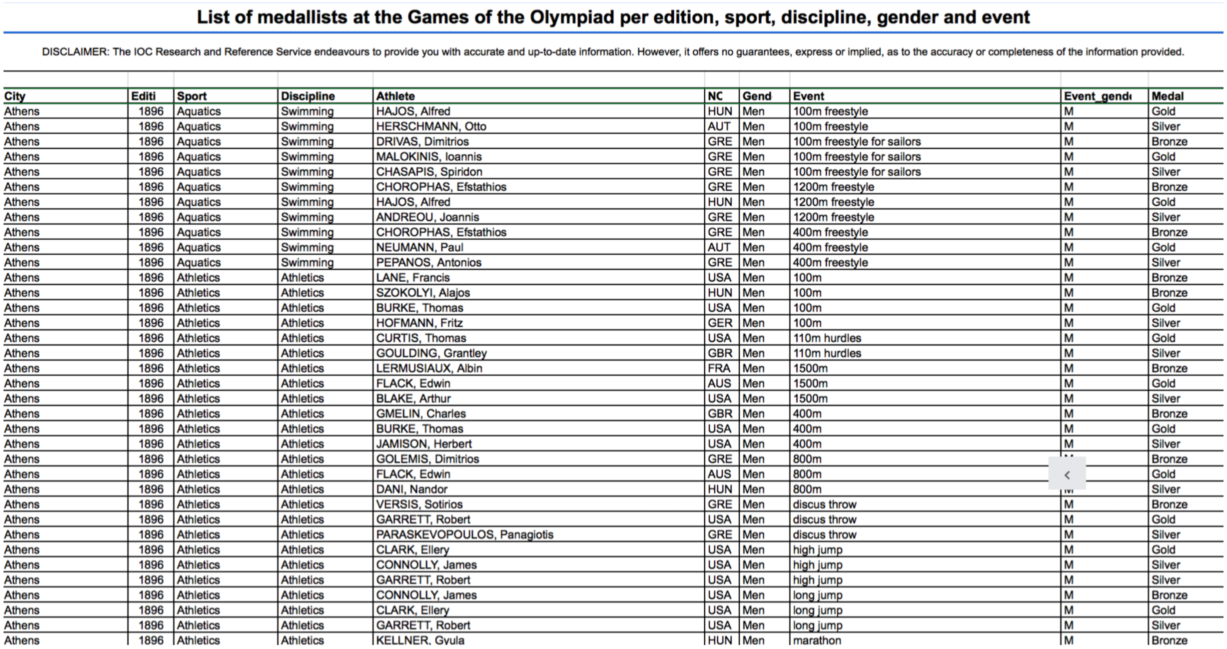
In our work we use a specific ontology[4] and have a different definition of data cleaning and cleansing. Our focus is on the axioms that previously have been defined in our ontology-base and will detect incompatibility among the records.

Methods

Dataset

In this paper, we are working on the summer Olympic medalists from 1896 to 2008 information. As you can see in Figure 1, the data consist of some features where the column with Medal header is important for us because we are testing our approach with multivariate analysis by SVM classification algorithm. The labels are Bronze, Silver, and Gold.

[1] Dirty data here means the data that need to be cleaned and cleansed and consist of irrelevant and incorrect records



List of medallists at the Games of the Olympiad per edition, sport, discipline, gender and event

DISCLAIMER: The IOC Research and Reference Service endeavours to provide you with accurate and up-to-date information. However, it offers no guarantees, express or implied, as to the accuracy or completeness of the information provided.

City	Editi	Sport	Discipline	Athlete	NC	Gend	Event	Event_gend	Medal
Athens	1896	Aquatics	Swimming	HAJOS, Alfred	HUN	Men	100m freestyle	M	Gold
Athens	1896	Aquatics	Swimming	HERSCHMANN, Otto	AUT	Men	100m freestyle	M	Silver
Athens	1896	Aquatics	Swimming	DRIVAS, Dimitrios	GRE	Men	100m freestyle for sailors	M	Bronze
Athens	1896	Aquatics	Swimming	MALOKINIS, Ioannis	GRE	Men	100m freestyle for sailors	M	Gold
Athens	1896	Aquatics	Swimming	CHASAPIS, Spiridon	GRE	Men	100m freestyle for sailors	M	Silver
Athens	1896	Aquatics	Swimming	CHOROPHAS, Efstathios	GRE	Men	1200m freestyle	M	Bronze
Athens	1896	Aquatics	Swimming	HAJOS, Alfred	HUN	Men	1200m freestyle	M	Gold
Athens	1896	Aquatics	Swimming	ANDREOU, Joannis	GRE	Men	1200m freestyle	M	Silver
Athens	1896	Aquatics	Swimming	CHOROPHAS, Efstathios	GRE	Men	400m freestyle	M	Bronze
Athens	1896	Aquatics	Swimming	NEUMANN, Paul	AUT	Men	400m freestyle	M	Gold
Athens	1896	Aquatics	Swimming	PEPANOS, Antonios	GRE	Men	400m freestyle	M	Silver
Athens	1896	Athletics	Athletics	LANE, Francis	USA	Men	100m	M	Bronze
Athens	1896	Athletics	Athletics	SZOKOLYI, Alajos	HUN	Men	100m	M	Bronze
Athens	1896	Athletics	Athletics	BURKE, Thomas	USA	Men	100m	M	Gold
Athens	1896	Athletics	Athletics	HOFMANN, Fritz	GER	Men	100m	M	Silver
Athens	1896	Athletics	Athletics	CURTIS, Thomas	USA	Men	110m hurdles	M	Gold
Athens	1896	Athletics	Athletics	GOULDING, Grantley	GBR	Men	110m hurdles	M	Silver
Athens	1896	Athletics	Athletics	LERMUSIAUX, Albin	FRA	Men	1500m	M	Bronze
Athens	1896	Athletics	Athletics	FLACK, Edwin	AUS	Men	1500m	M	Gold
Athens	1896	Athletics	Athletics	BLAKE, Arthur	USA	Men	1500m	M	Silver
Athens	1896	Athletics	Athletics	GMELIN, Charles	GBR	Men	400m	M	Bronze
Athens	1896	Athletics	Athletics	BURKE, Thomas	USA	Men	400m	M	Gold
Athens	1896	Athletics	Athletics	JAMISON, Herbert	USA	Men	400m	M	Silver
Athens	1896	Athletics	Athletics	GOLEMIS, Dimitrios	GRE	Men	800m	M	Bronze
Athens	1896	Athletics	Athletics	FLACK, Edwin	AUS	Men	800m	[illegible]	Gold
Athens	1896	Athletics	Athletics	DANI, Nandor	HUN	Men	800m	M	Silver
Athens	1896	Athletics	Athletics	VERSIS, Sotirios	GRE	Men	discus throw	M	Bronze
Athens	1896	Athletics	Athletics	GARRETT, Robert	USA	Men	discus throw	M	Gold
Athens	1896	Athletics	Athletics	PARASKEVOPOULOS, Panagiotis	GRE	Men	discus throw	M	Silver
Athens	1896	Athletics	Athletics	CLARK, Ellery	USA	Men	high jump	M	Gold
Athens	1896	Athletics	Athletics	CONNOLLY, James	USA	Men	high jump	M	Silver
Athens	1896	Athletics	Athletics	GARRETT, Robert	USA	Men	high jump	M	Silver
Athens	1896	Athletics	Athletics	CONNOLLY, James	USA	Men	long jump	M	Bronze
Athens	1896	Athletics	Athletics	CLARK, Ellery	USA	Men	long jump	M	Gold
Athens	1896	Athletics	Athletics	GARRETT, Robert	USA	Men	long jump	M	Silver
Athens	1896	Athletics	Athletics	KELLNER, Gyula	HUN	Men	marathon	M	Bronze

Figure1. This figure shows some records of our data that was used for this study

Implementing Points

In this study, we designed an ontology-based according to COGNIBASE with Protege software. After designing our ontology-based, we linked it to a python code by the means of Owlready module in python3. Actually, we put OWL/XML file of designed ontology-based in the directory of project Python code and used Owlready for loading it. Finally, we had both ontology-based and python codes in a file for editing and adding extra procedures to them. The most important part of ontology-based was axioms which help us to clean and cleans related data so we added some functions for detecting and modifying records that don’t meet axioms roles defined for ontology-based.

For testing our approach, we splitted 9,200 records for test and 20,000 records for training phase from 29,200 records in the main table. We considered these 20,000 records as actual dataset and made a dirty dataset from it with explained algorithm and make them clean and cleans(we call it cleaned dataset) with axioms were defined in the axiom class. We trained three SVM models for these three datasets(actual, dirty and cleansed) and tested them with 9,200 records. All these steps have been demonstrated in Figure 4.

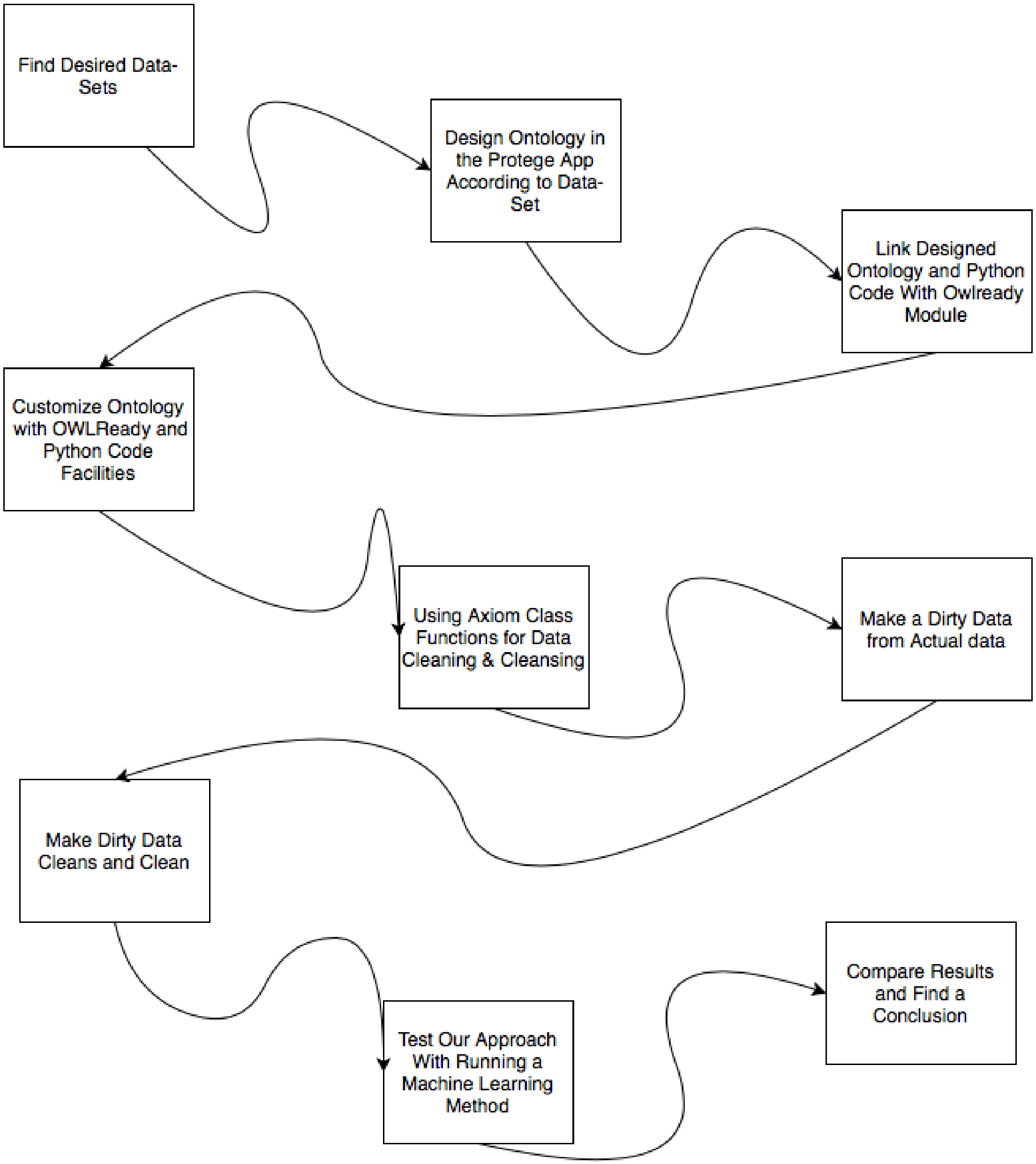

Find Desired Data-Sets
Design Ontology in the Protege App According to Data-Set
Link Designed Ontology and Python Code With Owlready Module
Customize Ontology with OWLReady and Python Code Facilities
Using Axiom Class Functions for Data Cleaning & Cleansing
Make a Dirty Data from Actual data
Make Dirty Data Cleans and Clean
Test Our Approach With Running a Machine Learning Method
Compare Results and Find a Conclusion

Figure 4. The general steps for this study are mentioned in this figure.

Results

Three vectors with 9,200 records were predicted from three SVM models of actual, dirty and cleansed datasets. We computed RMSE between actual and dirty predicted vector that was 0.48, in addition, we computed RMSE between actual and cleansed predicted vector that was 0.01. This difference between errors indicates the impact of our approach for data cleaning and cleansing process. You can see in figure 5 that the similarity between predicted vector for actual and cleansed data is much more than actual and dirty data.

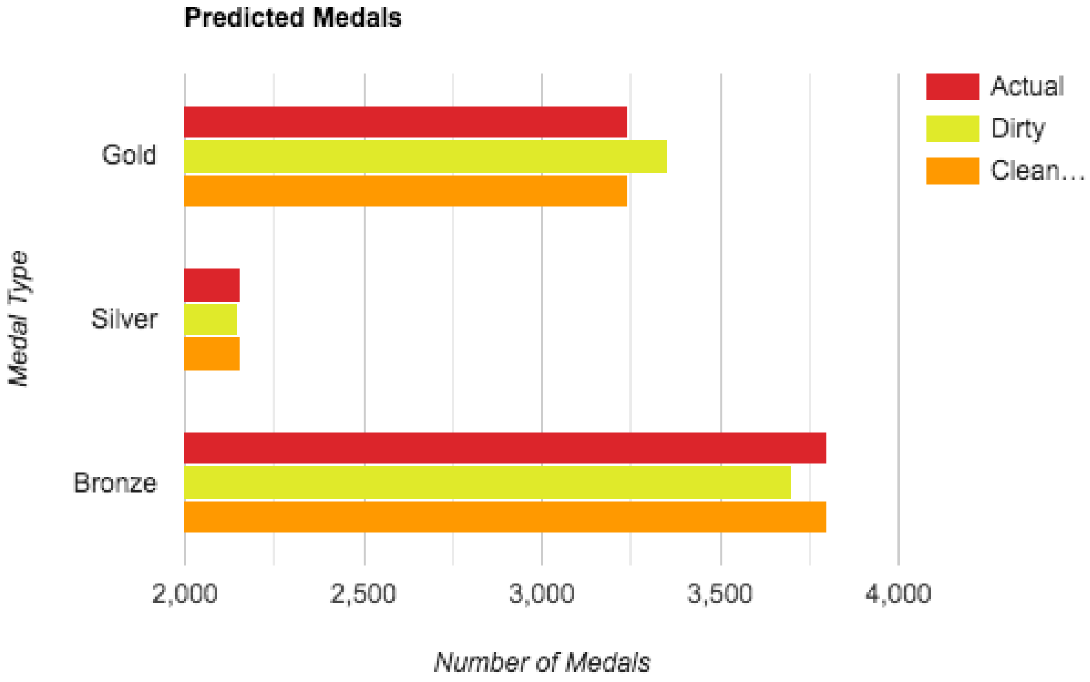


Figure 5. Predicted medals with three classifiers obtained from actual, dirty and cleansed datasets.

Designed Ontology-Based

The ontology-based was designed for summer Olympic games so it supports concepts that are common in Olympic games. As you can see in figure 2 the class diagram for our Olympic ontology-based consist of some concepts and their object properties. Figure 3 also shows the graph diagram of the designed Olympic ontology.

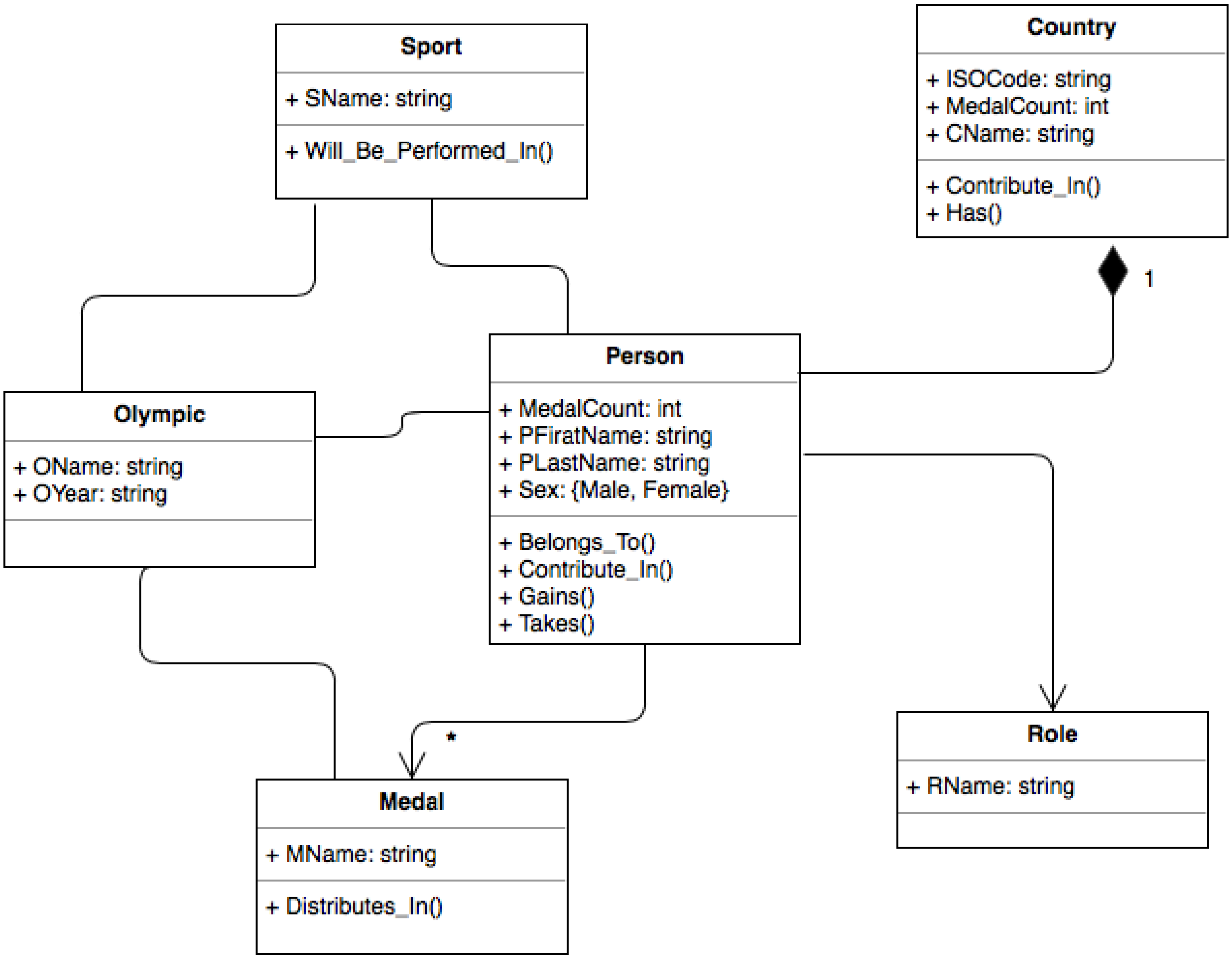


Figure 2. The class diagram for Olympic ontology-based

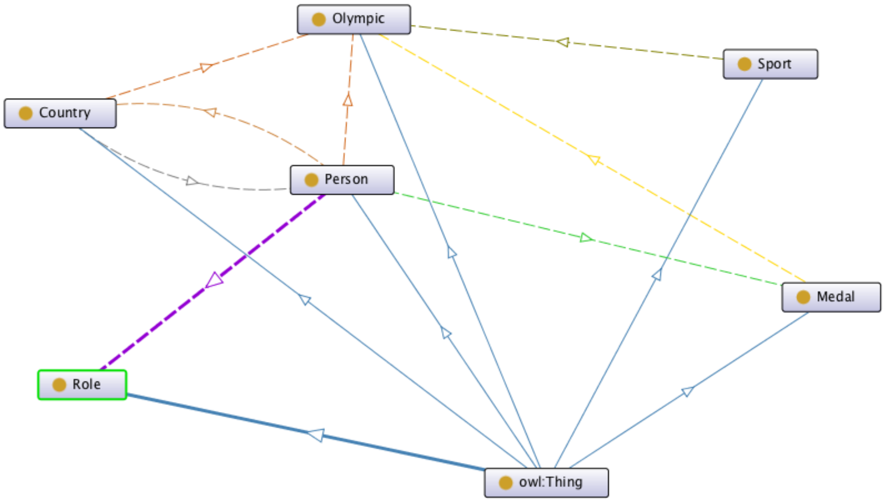


figure 3. The graph of Olympic ontology-based

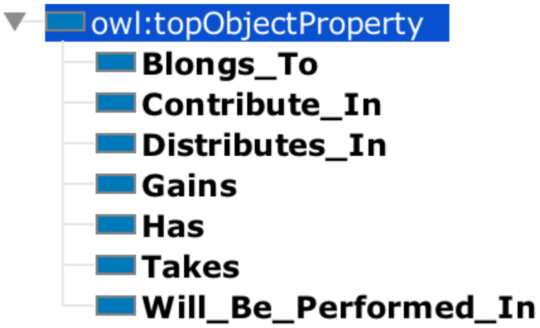


Figure 5. Object properties designed for our ontology-based by Protege.

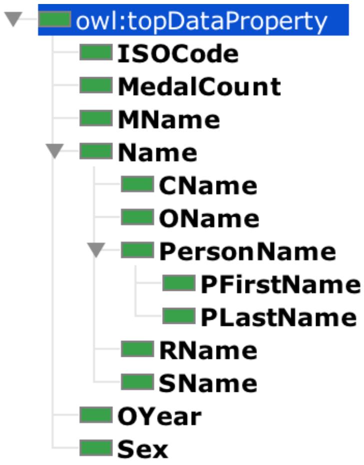
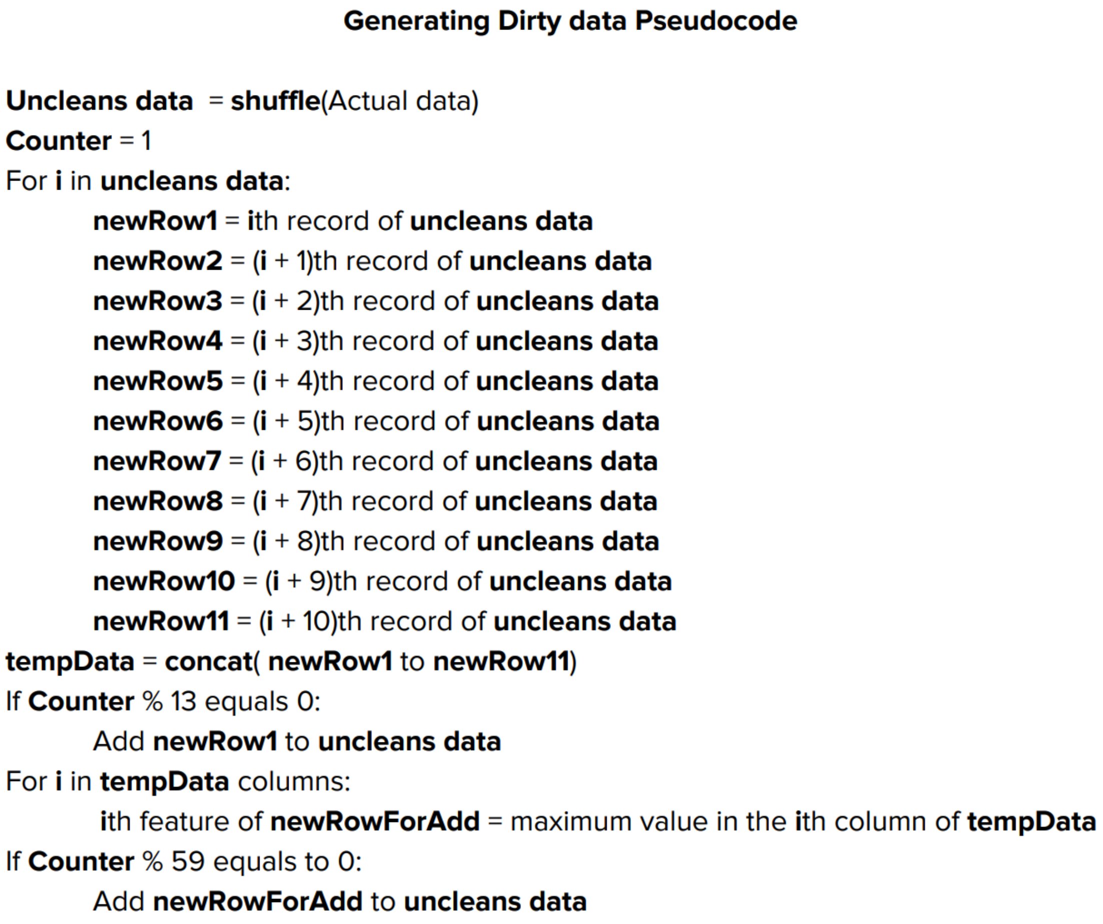


Figure 6. Data properties designed for our ontology-based by Protege.

Algorithm for producing dirty data

For generating data that are not compatible with Olympic medalist records we designed an algorithm that put every ten records in one group and selects the mean value of each feature for generating a new record. We iterate this procedure for all records and at the end, we have a lot of data that syntactically are true but don't meet our axiom limitations. For example, consider an athlete wherein one record we specified woman for its gender value and in another one we have a man for its gender value. Firstly, we should detect this incompatibility and secondly, we should correct it.



Generating Dirty data Pseudocode

```
Uncleans data  = shuffle(Actual data)
Counter = 1
For i in uncleans data:
        newRow1 = ith record of uncleans data
        newRow2 = (i + 1)th record of uncleans data
        newRow3 = (i + 2)th record of uncleans data
        newRow4 = (i + 3)th record of uncleans data
        newRow5 = (i + 4)th record of uncleans data
        newRow6 = (i + 5)th record of uncleans data
        newRow7 = (i + 6)th record of uncleans data
        newRow8 = (i + 7)th record of uncleans data
        newRow9 = (i + 8)th record of uncleans data
        newRow10 = (i + 9)th record of uncleans data
        newRow11 = (i + 10)th record of uncleans data
tempData = concat( newRow1 to newRow11)
If Counter % 13 equals 0:
        Add newRow1 to uncleans data
For i in tempData columns:
        ith feature of newRowForAdd = maximum value in the ith column of tempData
If Counter % 59 equals to 0:
        Add newRowForAdd to uncleans data
```

Figure 7. The pseudocode of the designed algorithm to make an uncleansed dataset from the actual dataset. According to this procedure, we made a dirty dataset from summer Olympic medalist dataset. We made a clean and cleansed dataset from the output of this algorithm in the next phase and called it cleansed dataset.

Axioms for Data Cleaning and Cleansing

As we know, axioms define specific roles for individual records that help us know more details about the domain and remove incompatible ones that don't meet the limitations. In this study, we defined some functions for axiom class that we have access to change it via Owlready module in Python. Gender-Check, Major-Check and Monality-Check are functions that detect records which don't meet axiom class limitation and are incompatible. Gender-Modify, Major-Modify, Monality-Modify and Clean-Dup are functions that modify detected records with correct value or remove them. For more details about having each limitation was defined and how we use it for detecting and modifying records, the source code is freely available at https://github.com/morteza-mahdiani/B.Sc-Thesis.

With the explained procedure and algorithm used in source code, we will make clean and cleans data from dirty data. Actually, we used axiom class of designed ontology-based to improve the cleaning and cleansing procedure, not only in finding syntactically incorrect records but also in finding semantically incorrect ones.

Conclusion

Finally, we could implement a method for data cleaning and cleansing with an ontology-based approach. The results show a meaningful difference between errors for dirty and cleansed data and actual ones and demonstrate that this method can be used for preprocessing phase of data analyzing.

Discussion

Using an ontology-based for data cleaning and cleansing has some advantages and we can detect incompatible records in both syntax and semantic level but there are still some difficulties that make this procedure a bit hard to implement. Difficulties like designing an ontology that needs an expert especially for defining axioms. We should find a way to do this in the future. Maybe ontology learning methods will be useful for designing these axioms in the future.

Acknowledgement

I have to thank Prof. Abdollah-Zadeh Barforoush for his tips and help during doing this study. I appreciate Ms Samin Jamshidi for his helpful guides during implementing the code and Mr Amir-Ahmad Habibi for his helpful comments.

References

[1]B. Boehm, “Software Engineering: R & D Trends and Defence Needs”, Chapter 19 in Research Direction in Software Technology (P. Wegner ed.), Cambridge, MA, MIT Press, 1979.

[2] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. Mach. Learn. 20, 3 (September 1995), 273-297.

[3] A. A. Barforush and A. Rahnama. Ontology learning: Revisted. Journal of Web Engineering, 11(4):269–289, May 2012. URL https://dl.acm.org/citation.cfm?id=2481557.

[4] A. Rahnama and A. A. Barforush. Cognibase: A new representation model to support ontology development. IADIS International Conference Information Systems, 243–248, - 2011.

[5] Paulo Oliveira, Fátima Rodrigues, and Pedro Henriques. An ontology-based approach for data cleaning. Proceedings of the 11th International Conference on Information Quality, November 2006.

[6] A. S. Arnold, J. S. Wilson, and M. G. Boshier. A simple extended-cavity diode laser. Review of Scientific Instruments, 69(3):1236–1239, March 1998. URL http://link.aip.org/link/?RSI/69/1236/1.

[7] N. F. Noy and D. L. McGuinness. Ontology Development 101: A Guide to Create Your First Ontology

[8] X. Wang, J. Hamilton and Y. Bither. An Ontology-Based Approach To Data Cleaning. July, 2005.

[9] Z. Khedad and E. Matais. Ontology-Based Data Cleaning. 2002

[10] C. Batini, C. Cappiello, C. Francalanci and A. Maurino. Methodologies for Data Quality Assessment and Improvement. July, 2009.

[11] Summer Olympic Medalist 1986 to 2008, URL: www.theguardian.com

[12] C.Cortes and V.Vapnik. Support-Vector Networks, Machine Learning, 273-297, 1995.

[13] Rational Unified Process white paper, Archived, 2009-05-01, at the Wayback Machine.

[14] Hyndman, Rob J., Koehler, Anne B. "Another look at measures of forecast accuracy". International Journal of Forecasting, 22 (4): 679–688, 2006.

[15](2007) Ontology in Computer Science. In: Semantic Web: Concepts, Technologies and Applications. NASA Monographs in Systems and Software Engineering. Springer, London