AUTOMATIC VARIABLE DETECTION AND FORMATTING FOR CROSS-DISCIPLINARY DATA SET COMPATIBILITY

Alexander Moreno[1], Keval Shah[2], and Yuan Zhao[3]
[1] Department of Computer Science, Georgia Institute of Technology, Atlanta, GA 30332; [2] Department of Analytics, University of Chicago, Chicago, IL 60637; [3] Department of Bioinformatics, University of California, San Diego, La Jolla, CA 92038

ABSTRACT

Data sets from disparate sources often employ different formatting styles for the same variables. We propose the development of a tool for automatic detection of shared parameters and conversion into a common, compatible format. This tool will support a large number of commonly measured variables, including timestamps and geographical location. As a proof of concept, we implemented this tool for unifying date-time information in csv files. Finally, we plan to enable automatic detection of dataset variables and use this information to build a cross-compatibility database of shared parameters across datasets from disparate repositories, such as OSDC and data.gov.

1 INTRODUCTION

Fundamentally, any two data sets can be compared and co-analyzed given that they share at least one common parameter; however, data from varied sources often use different formatting styles and to analyze these data with one another, it is necessary to first process the information into a compatible format. This task is not necessarily difficult to perform depending on the variables involved, but the required formatting can be a tedious and inefficient use of time.

To solve this data processing problem, we propose the development of a tool for automatic parsing of data variables and formatting them for compatibility. This tool will support a large number of commonly measured variables, including but not limited to timestamps, geographical location, distances, and weights for the purpose of classification. The tool's supported parameters list can be extended with plugins, so support for additional variables can added in the future, including uncommon measurements for specific experiments. We also plan to implement a function that will automatically screen data sets from various data sources, including OSDC and data.gov, enumerate the variables present in each data set, and export a list of shared variables between data sets of interest. This tool will ultimately aid in facilitating data intensive cross-disciplinary discoveries by reducing programmer time lost to data processing while also identifying common dimensions for analysis across data sets.

2 METHODS

Our tool has two main capabilities: (1) data format parsing and conversion and (2) shared parameter extraction from multiple datasets.

The first function can be broadly described in four steps. For each variable (column), the column header will be read to determine the semantic meaning of the header string. This string will be cross-referenced with word clusters for each variable to decide which variable the header represents; for instance, the time parameter can be represented by 'time', 's', and 'hrs', among others. Next, the entries will be parsed into an object for manipulation. Based on analysis of the format and determination of the variable, appropriate logical rules and conventions will be applied and the entries will be re-arranged.

The second function will take as its input at least two data sets and output an 'associated variable' dictionary. For each dataset, its column headers will be enumerated and added to the dictionary in the format of *variable:[list of datasets]*, where the variable represented by column (determined by the same word clustering method described above) is the key and the dataset name is the value. From this dictionary, it is trivial to see which variables are present in the superset of data sets, which data sets share which variables for co-analysis, and how prevalent each variable is across the data sets.

3 CHALLENGES

Examples of some challenges could include delimiters, abnormal formatting, meta data descriptors (i.e. super structured data where relevant information such as dates or experiment are present in the file names or parent folders), and the missing data problem. For the missing data problem, one solution would be to detect

whether the data is time series data and filling empty entries with NA/NaN, or employing a keyword argument that defaults to row deletion if the data is not time series. These strategies are flexible and other fill options may be used as necessary for the data type in question.

Overall, these challenges relate to the larger problem of being able to handle many formats. Due to the limitless number of potential variables, it is intractable to write formatting code for every single one. However, it is possible to write plugins for the majority of commonly measured variables, which will be discussed further below.

4 RESULTS & DISCUSSION

As a proof of concept, we implemented a preliminary build of the tool, which is available at https://github.com/onenoc/OSDCcompetition. We anticipate that dates and time will be one of the most commonly measured variables, so for this build, we implemented the parsing of date strings and conversion into datetime objects as the proof for the formatting function. We applied this function to two independent datasets, twitter data and weather in Chicago, which could not be easily analyzed together in their raw forms due to differences in the format as well as gaps in sampling time.

Further, to illustrate our approach toward determining the prevalence of variables across data sets, we generated the following word cloud with simulated variables and frequencies, where color represents frequency (gray means higher prevalence), though other visual aspects such as word size can also be used. As mentioned previously, we expected dates and time to be the most commonly occurring column name, with others such as city and location also being frequently measured variables. This visualization captures one of the types of relationships that can be discovered from our proposed tool's second function, and we plan to use this method in practice on OSDC data sets and those from other repositories to prioritize writing formatting code for variables.

Figure 1. Simulated word representing frequency of variables in several datasets. Gray corresponds to higher frequency.

Ultimately, this tool will facilitate cross-disciplinary data intensive analysis by enabling researchers to quickly identify shared variables in datasets of interest and convert their values to a common format for faster analysis. Furthermore, this tool will allow for the extraction of shared variables between any number of data sets, in addition to providing information on variable frequency, as shown by the word cloud. Last, this tool can encourage the use of partner sites' data through the ability to identify data sets that share variables, which can potentially lead to novel and unconventional data comparisons, such as weather and crime rates.

5 ACKNOWLEDGEMENTS

We acknowledge Dr. Margaret Bernard, Dr. Isao Kojima, and Dr. Maria Patterson for advice and guidance during the project idea formulation. We also acknowledge Jeffrey Weekley for the further guidance, idea development, and visualization. Finally, we wish to acknowledge Dr. Bob Grossman and Dr. Heidi Alvarez, as well as the University of Amsterdam faculty, for directing and hosting the 2014 OSDC PIRE Workshop.

6 REFERENCES

[1] Wunderground www.wunderground.com
[2] Scraper wiki https://scraperwiki.com/
[3] Twitter API documentation https://dev.twitter.com/docs