UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

COS700
Research Proposal

Data Pre-Processing Framework

Student number: 12230830

Supervisor(s):
Will van Heerden

2018/05/01

Data Pre-Processing Framework

Abstract

Data mining deals with the extraction of knowledge from large amounts of data[HK06]. A key portion of this process is the data preparation portion of this data mining phases. This portion of phases includes data cleaning, data integration, data selection and data transformation. These are the first four steps in the knowledge discovery process. There are multiple techniques involved in each of these phases and the entire data mining process as a whole is quite time-consuming. Data comes in all forms and sizes. Furthermore, multiple sources of data exist and all require some level of pre-processing before any information can be derived from the data. This paper proposes a software solution in the form a data pre-processing framework which will facilitate the creation of a connected data flow, starting with one or more source(s) in different formats. Data preparation is to be performed on this data until it reaches a 'ready' state after which further processing may occur.

Keywords:

data, mining, pre-processing, framework, reduction, scaling, cleaning, windowing, consolidate

1 Introduction

As technology advances and basic human tasks are simplified in order to increase productivity and efficiency in general, we become overwhelmed with data; financial data, demographic data and data from social media platforms. The ability to manually gather and analyse all of it has transcended human capacity and we have therefore arrived at a need to automatically acquire, process, understand and express information from this data[HK06]. Knowledge discovery has become an integral part of our work. We especially see this in areas such as digital forensics, fraud detection, consumer trends and market analysis, for example. Data mining, a fairly young discipline deals with the extraction of knowledge from large amounts of data.

Knowledge Discovery from Data is often synonymously coupled with data mining. It is a process consisting of different steps aimed at mining and understanding knowledge from these huge datasets. The steps include data cleaning, data integration, data selection and data transformation. These are the first four steps in the knowledge discovery process.[HK06] They are also the variant categories of Data Pre-processing.

Data comes in all kinds of shapes and sizes. This ranges from numeric data to graphic and chart data. Furthermore, data sources generally contain noisy, incomplete and bad data. It is therefore a requirement to support data from different kinds of sources as well the format in which the data may come. We need to process said data into meaningful and useful data to increase the speed, reliability, quality and cost of the entire data mining process. The process of applying techniques to achieve this is called Data Pre-processing[HK06][Nor16]. We are tasked with employing these techniques in an elegant and scalable fashion.

"The objective is to develop a generic software framework which provides a mechanism for creating a connected data flow, starting with one or more original and unprepared data sets (which may appear in a variety of formats), through a series of preparation tasks, until a single cleaned dataset is produced." This paper outlines some important aspects and steps to be followed in the planning, design and implementation of this software framework. In the next section we discuss the problem statement in greater detail. The Literature Study in the section afterwards will discuss, in depth the findings from articles and publications with the data mining field in general as well as within the data pre-processing area.

2 Problem Statement

Let us consider a very special scenario. Imagine that you are very skilled data analyst at a huge investment company. Now, throughout the years, a wide range of transactions have occurred within the company's systems and you are tasked with gathering and interpreting empirical data from each of these transactions. You are required to identify trends, attributes and key indicators accurately and give a full report of your findings. The first problem you encounter is the need to gather the data from the company's wide range of database management sys-

tems; SQL Server, Oracle, PL/SQL and a recent MongoDB installation. This hurdle can quite inefficiently be conquered to produce the required data.
Upon the tedious consolidation of these different kinds of data, you start to notice that some attributes had not been captured or had been omitted. Some of the attributes have values not pertaining to the company standards âĂŞ they are abnormal. Some data also differ in categorizations from data source to data source (e.g., withdrawals have a category ID of 'wd' in SQL Server-derived data and 'wdl' in Oracle).

In light of the need for a faster, robust data pre-processing framework, we arrive at the following research question:

Is it feasible to design and implement a generic, modifiable data pre-processing framework intelligent enough to support cross-platform integration with different, unprepared data sources to produce a single cleaned data set?

The main objective is to answer the following sub-questions:

I Can we have a software framework generic enough to form part of a larger library of data mining techniques?

II Given multiple sources and formats of data, how can we integrate and automate the data extraction process from these sources?

III What are the existing data mining frameworks and how does our design improve on the non-functional requirements of the data mining process with special attention to speed?

3 Literature Study

3.1 Data Pre-processing

Data tend to be inconsistent, incomplete and dirty. It may take a long period of time to spot these anomalies between data and values to attributes may need to be derived to complete the data. Data pre-processing helps to improve the quality of the data. With complete, consistent and clean data, we can facilitate a more successful knowledge discovery process. Remember, data pre-processing only forms part of the longer data mining chain of actions. Data preparation techniques help us to arrive at more accurate results through a more efficient process[HK06][FPSS94].

3.1.1 Data Cleaning

In attempt to solve the problem of real-world data be inconsistent, noisy and inconsistent, data cleaning routines try to consolidate any problems with the data; uncaptured values, outliers etc. As part of the process, data cleaning detects discrepancies, noise and missing values. A series of steps is performed and metadata are carefully considered based on pre-existing knowledge.

3.1.2 Data Integration and Transformation

Data Integration is one of the keystones of the proposed software framework. One prevalent task in data preparation is gathering data from multiple sources and consolidating it into a single data store. This is a very trick problem to have because of the differences in multiple data stores such as the format the data which are stored, the way in which the data are queried and schemas and general data organisation. Data can come from relational databases, object-relational databases, csv files, text files, NoSQL databases etc. to name a few. Data integration is necessary because we want the ability to integrate with these different sources and formats to be supported by our data pre-processing software framework[FPSS96][FPSS94].

Data Transformation refers to the process that comes after all data has been gathered and considered. The data are transformed, refined and translated into a standard, consolidated form which is ready to be mined. Our software framework has to incorporate a way to employ data transformation processes such as smoothing, aggregation and normalization - which will be explored in more technical detail in our research paper. Data transformation is very the necessary for the purpose of this research. Because we are expecting different data from different sources, part of the automated process will be to carefully and intelligently transform this data into a uniform, consolidated state.

3.1.3 Data Reduction

The term 'Data Reduction' seems quite self-explanatory. It refers to techniques used to an abstracted, reduced representation of data. As one can imagine, retrieving millions of records from an investment company's database could potentially return an insolubly large amount of data[HK06]. We therefore need to employ data reduction techniques such as data cube aggregation, attribute subset selection and numerosity reduction[BRB93]. These techniques will help us efficiently arrive at a subset (or superset) of data which can be relatively easily processed. Our data pre-processing software framework needs to accommodate this functionality by employing clever ways of deriving important information from attributes.

3.2 Tying it all together

With each of these major data pre-processing techniques in place, we aim to arrive at an abstract, robust and efficient software solution that facilitates the data pre-processing sequence of events. This software framework will be quite relevant in the further exploration of simplifications of data mining.

4 Methodology

4.1 Approach

4.1.1 Literature Surveys

Because our research focus is one directly targeted at using existing concepts and tying them down into a software solution framework, literature surveys of existing publications will be very necessary.

4.1.2 Prototype: Proof-of-concept

A proof-of-concept prototype showing that the software solution can be successfully and usefully implemented will be done.

4.1.3 Experiments

Experiments will be run using this prototype to illustrate the compliance with non-functional requirements. A series of simulations will be run throughout the research process to identify key attributes of our research e.g., how easily SQL and NoSQL databases can integrate with each other.

5 Planning

Attached to the document is the Project Gantt chart. A PDF document of the chart and be found at: https://bit.ly/2HVk7dl

At Risk	Task Name	Status	Start Date	End Date	% Compl...	Notes
	Phase 1: Inception		**05/04/18**	**10/06/18**	**0%**	
	Gather all information concerning existing frameworks	In Progress	05/04/18	05/11/18	10%	
	Write out functional and non-functional requirements	Not Started	05/11/18	05/12/18	0%	
	Document inception findings	Not Started	05/04/18	10/06/18	0%	
	Research data preparation	Not Started	05/12/18	05/31/18	0%	
	Phase 2: Prototype development		**06/01/18**	**08/25/18**		
	Backlog creation and prioritization	Not Started	06/01/18	06/06/18	0%	
	Sprint planning	Not Started	06/06/18	06/08/18	0%	
	Data source integration tool (dev)	Not Started	06/09/18	06/23/18	0%	
	Data pre-processing specifications	Not Started	06/24/18	06/29/18	0%	
	Data cleaning tool	Not Started	06/30/18	07/28/18	0%	
	Data transformation tool	Not Started	07/30/18	08/10/18	0%	
	Data reduction platform	Not Started	08/11/18	08/25/18	0%	
	Phase 3: Research documentation and analysis		**08/26/18**	**10/06/18**		
	Software documentation	Not Started	08/26/18	09/07/18	0%	
	Hardening: bugfixes and deployments	Not Started	09/07/18	09/21/18	0%	
	Research question evaluation	Not Started	09/22/18	10/06/18	0%	
	Did we answer our research question?					
	Have we answered all sub-questions?					

References

[BRB93] Philip Bevington, Keith Robinson, and Gerry Bunce. Data reduction and error analysis for the physical sciences. *American Journal of Physics*, 61(8):766–766, 1993.

[FPSS94] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. *Knowledge Discovery and Data Mining*. AAAI Press, Seattle, WA, 1994.

[FPSS96] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. Towards a unifying framework. In *Knowledge Discovery and Data Mining*, 1996.

[HK06] Jiawei Han and Micheline Kamber. *Date Mining Concepts and Techniques*. Morgan Kaufmann Publlishers, Reading, Massachusetts, 2 edition, 2006.

[Nor16] Matthew North. *Data Mining For the Masses With Implementations in RapidMiner and R*. Middleton DE, USA, Middleton, 2 edition, 2016.