CerFix: A System for Cleaning Data with Certain Fixes

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Early version, also known as pre-print

Published In:
Proceedings of the VLDB Endowment (PVLDB)

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
CerFix: A System for Cleaning Data with Certain Fixes

Wenfei Fan 1,2 Jianzhong Li2 Shuai Ma3 Nan Tang1 Wenyuan Yu1
1University of Edinburgh 2Harbin Institute of Technology 3Beihang University
{wenfei@inf., ntang@inf., wenyuan.yu}@ed.ac.uk lijzh@hit.edu.cn mashuai@act.buaa.edu.cn

Abstract

We present CerFix, a data cleaning system that finds certain fixes for tuples at the point of data entry, i.e., fixes that are guaranteed correct. It is based on master data, editing rules and certain regions. Given some attributes of an input tuple that are validated (assured correct), editing rules tell us what other attributes to fix and how to correct them with master data. A certain region is a set of attributes that, if validated, warrant a certain fix for the entire tuple. We demonstrate the following facilities provided by CerFix: (1) a region finder to identify certain regions; (2) a data monitor to find certain fixes for input tuples, by guiding users to validate a minimal number of attributes; and (3) an auditing module to show what attributes are fixed and where the correct values come from.

1. Introduction

It has long been recognized that real-life data is often dirty [10]. Dirty data is costly: it costs US businesses 600 billion dollars each year [5]. This highlights the need for data cleaning tools, to help users detect errors in the data and moreover, repair the data, i.e., correct the errors.

Most data cleaning systems on the market are ETL tools (extraction, transformation, loading; see [8] for a survey). To detect semantic errors in the data, there have also been a host of approaches on data repairing based on integrity constraints [1, 2, 4, 11]. A variety of constraints have been studied for data repairing, such as traditional functional, inclusion and full dependencies [1, 2], as well as their extensions (e.g., conditional functional dependencies [4, 11]).

Integrity constraints are capable of detecting the presence of errors in the data, i.e., determining whether the data is dirty or not. However, they do not tell us which attributes of a tuple have errors and how we could correct the errors.

Example 1: Consider an input tuple:

\( t: (\text{FN} = \text{Bob}, \text{LN} = \text{Brady}, \text{AC} = 020, \text{phn} = 079172485, \text{type} = 2, \text{str} = 501 \text{ Elm St}, \text{city} = \text{Edi}, \text{zip} = \text{EH8 4AH}, \text{item} = \text{CD}) \)

It specifies a UK customer: name (\( \text{FN}, \text{LN} \)), phone number (area code AC and phone phn), address (street str, city, zip code) and item purchased. Here phn is either home phone or mobile phone, indicated by type (1 or 2, respectively).

To specify the semantics of the input data, one may define the following conditional functional dependencies (cfd):

\( \psi_1: \text{AC} = 020 \rightarrow \text{city} = \text{Ldn}, \quad \psi_2: \text{AC} = 131 \rightarrow \text{city} = \text{Edi} \).

These cfd statements that if AC is 020, then city should be Ldn (\( \psi_1 \)), and when AC is 131, then city must be Edi (\( \psi_2 \)). They detect the presence of errors in tuple \( t[\text{AC}] = 020 \) but \( t[\text{city}] = \text{Edi} \). However, they do not tell us which of \( t[\text{AC}] \) and \( t[\text{city}] \) is wrong, and to what value it should be changed.

In light of this, previous constraint-based methods use heuristics: they do not guarantee correct fixes in data repairing. Worse still, they may introduce new errors when trying to repair the data. Indeed, the correct values of \( t[\text{AC}], \text{city} \) are (131, Edi). Indeed, all these previous methods may opt to change \( t[\text{city}] \) to Ldn; this does not fix the erroneous \( t[\text{AC}] \) and worse, messes up the correct attribute \( t[\text{city}] \).

In practice, one often wants to find certain fixes for the data, i.e., fixes that are guaranteed to be correct. In other words, if a data repairing process changes the value of an attribute \( t[A] \) to \( v \), then \( v \) should be the true value of \( t[A] \). In addition, it should ensure that no new errors are introduced in the process. The need for certain fixes is particularly evident when repairing critical data, e.g., medical data, in which a seemingly minor error may mean life or death.

An approach to finding certain fixes was proposed in [7], based on master data, editing rules and certain regions. Master data (a.k.a. reference data) is a single repository of high-quality data that provides various applications in an enterprise with a synchronized, consistent view of its core business entities [9]. It is increasingly common for enterprises to maintain master data nowadays, and master data management (MDM) is being developed by IBM, SAP, Microsoft and Oracle. In contrast to integrity constraints, editing rules aim to tell us which attributes of a tuple are wrong and what values from master data they should take, provided that master data is available and that some other attributes are validated (assured correct). With respect to a set of editing rules and a master relation, a certain region is a set of attributes and a pattern tableau, such that for any input tuple \( t \), if \( t[A] \) is validated and \( t[A] \) bears a pattern in the tableau, then the editing rules guarantee to find a (unique) certain fix for all the attributes of \( t \).

Example 2: Suppose that master data contains a tuple:

\( s: (\text{FN} = \text{Robert}, \text{LN} = \text{Brady}, \text{AC} = 131, \text{Hphn} = 6884563, \text{Hphn} = 079172485, \text{str} = 501 \text{ Elm St}, \text{city} = \text{Edi}, \text{zip} = \text{EH8 4AH}, \text{DoB} = 11/11/55, \text{gender} = \text{M}) \)

An editing rule is given as follows:

\( \varphi_1: (\text{zip}, \text{zip}) \rightarrow (\text{AC}, \text{AC}), (\text{t}_{\varphi_1} = ()) \)

It states that for an input tuple \( t \) and a master tuple \( s \), if they have identical zip code (i.e., \( t[\text{zip}] = s[\text{zip}] \)), then \( t \) could be updated by \( t[\text{AC}] := s[\text{AC}] \), provided that \( t[\text{zip}] \) is
correct (either ensured by the users or verified automatically). When the master data is accurate, the editing rule is assured valid, and the region \( t[\text{zip}] \) is correct, the fix to \( t[\text{AC}] \) with the master data value is certainly correct.

Following [7], we develop CerFix, a data cleaning system that finds certain fixes for input tuples at the point of data entry. It differs from other systems that also ask for user feedback (e.g., [3]) in what feedback is requested and how the feedback is used. Below we first present CerFix, and then outline what functionalities we shall demonstrate.

2. The CerFix System

The architecture of CerFix is depicted in Fig. 1. CerFix maintains a collection of master data (master data manager) and a set of editing rules (rule engine). With respect to the master data and editing rules, it computes a set of certain regions (region finder). It inspects and repairs input tuples via a data monitor, which interacts with the users to find certain fixes for input tuples. It also provides a Web interface (data explorer) and a data auditing module for users to manage editing rules and trace changes to data, respectively.

Below we briefly present the key components of CerFix. We refer the interested readers to [7] for details about editing rules, certain regions and their computation.

Rule engine. It maintains a set of editing rules (eRs) that specify (a) whether an input tuple \( t \) can match a master tuple \( s \) via a pattern tuple (e.g., (zip, zip)) together with \( t[p_1] \) in rule \( \varphi_1 \) of Example 2), and (b) which attribute values of \( t \) can be changed and what correct values they should take from the master data. The engine also implements static analysis techniques developed in [7]. In particular, it supports the following. (1) It checks the consistency of editing rules, i.e., whether the given rules are dirty themselves. Moreover, (2) provided that some attributes of a tuple are correct, it automatically derives what other attributes can be validated (assured correct) by using editing rules and master data.

Editing rules can be either explicitly specified by the users, or derived from integrity constraints, e.g., CFDs and matching dependencies [6] for which discovery algorithms are already in place. CerFix currently only supports manual specification of editing rules via the Web interface.

Master data manager. It maintains master data, which is assumed consistent and accurate [9].

Region finder. A region is a pair \((Z, T_c)\), where \(Z\) is a list of attributes of an input tuple and \(T_c\) is a pattern tableau consisting of a set of pattern tuples with attributes in \(Z\).

A region \((Z, T_c)\) is a certain region w.r.t. a set of editing rules and master data if for any input tuple \(t\), as long as \(t[Z]\) is correct and \(t[Z]\) matches a pattern in \(T_c\), the editing rules warrant to find a certain fix for \(t\). Based on the algorithms in [7], top-\(k\) certain regions are pre-computed that are ranked ascendingly by the number of attributes, and are recommended to users as (initial) suggestions.

Data monitor. This is the most important module of CerFix. It interacts with the users and finds certain fixes for input tuples, as follows. (1) Initial suggestions. It recommends the set of certain regions computed by region finder to the users as suggestions. For each input tuple \(t\), if the users ensure that \(t[Z]\) is correct and matches a pattern in \(T_c\) for any region \((Z, T_c)\) in the set, then a certain repair for \(t\) is warranted. (2) Data repairing. For an input tuple \(t\), the users may respond with a set \(t[S]\) of attributes that is correct, where \(S\) may not be any of the certain regions. Data monitor iteratively employs editing rules and master data to fix as many attributes in \(t\) as possible, and expands the correct attribute set \(S\) by including those attributes that are validated via the inference system of the rule engine. (3) New suggestion. If not all attributes of \(t\) have been validated, data monitor computes a new suggestion, i.e., a minimal number of attributes, which are recommended to the users. If the users ensure the correctness of these attributes in \(t\), data monitor will find a certain fix for \(t\). The process of steps (2) and (3) repeats until a certain fix of \(t\) is reached.

CerFix ensures that each fix is correct with editing rules and master data. It also minimizes users’ effort by identifying a minimal number of attributes for users to validate.

Data auditing. This module keeps track of changes to each tuple, incurred either by the users or automatically by data monitor with editing rules and master data. Statistics about the changes can be retrieved upon users’ requests.
3. Demonstration Overview

We next describe various aspects of CerFix in more detail, and explain the aims of our demonstration. More specifically, we show the following: (1) how users manage editing rules with the aid of the Web interface (data explorer in Fig. 1); (2) how CerFix interacts with the users for data monitoring, to detect and fix errors in input tuples at the point of their entry; and (3) how data auditing works, to keep track of which attributes are fixed and where the correct values come from, and provide statistics about the percentage of data that is fixed by user efforts or by CerFix.

Initialization. The users are required to configure an instance, which consists of two parts: (a) a data connection with JDBC url, username, password and the corresponding JDBC driver provided by users; and (b) specifying the schema of input (dirty) tuples and that of the master data.

We illustrate these with the master data and the input data shown in Fig. 2. Note that they have different schemas.

Master data. Each tuple in the master data specifies a person in the UK in terms of the name (FN, LN), area code (AC), home phone (Hphn), mobile phone (Mphn), address (street str, city and zip code), date of birth (DOB) and gender.

Input tuples. Each tuple specifies a UK customer, as illustrated in Example 1.

Editing rule management. Figure 2 displays the Web interface for managing eRs. In the demonstration we shall show how the users can manage (view/modify/add/delete) eRs using the data explorer. The system currently only supports to import eRs manually via the rule manager, where the eRs may either be designed by experts or be discovered from CFDs or MDs. For instance, Figure 2 shows nine editing rules \( \varphi_1 \), \( \varphi_2 \) and \( \varphi_3 \), for the id’s 1–9, respectively.

- \( \varphi_1 \) (resp. \( \varphi_2 \) and \( \varphi_3 \)) states that if an input tuple \( t \) and a master tuple \( s \) have the same zip code and if \( t[zip] \) is already validated, then \( t \) can be updated by \( t[zip] := s[zip] \) (resp. \( str \) and \( city \)),
- \( \varphi_2 \) (resp. \( \varphi_3 \)) states that if the \( \text{phn} \) of a tuple \( t \) matches the \( \text{Mphn} \) of a master tuple \( s \), and if \( t[\text{phn}] \) is validated, then \( t[FN] := s[FN] \) (resp. \( LN \)). These eRs pose a constraint (a pattern tuple) \( t[\text{type}] = 2 \), requiring that \( \text{phn} \) is mobile phone. This rule can be viewed or edited by clicking the view/edit frame for the pattern.
- \( \varphi_3 \) (resp. \( \varphi_7 \) and \( \varphi_8 \)) tells us that if the (AC, \( \text{phn} \)) attributes of an input tuple \( t \) match the (AC, Hphn) values of a master tuple \( s \), and if \( t[\text{AC}, \text{phn}] \) are validated, then \( t[\text{str}] := s[\text{str}] \) (resp. \( \text{city} \) and \( \text{zip} \)). These eRs have a pattern \( t[\text{type}] = 1 \), i.e., \( \text{phn} \) is home phone.
- \( \varphi_3 \) states that when the AC value of an input tuple \( t \) is not 0800 (toll free, non-geographic), if it agrees with a master tuple \( s \) on its AC attribute, and moreover, if \( t[\text{AC}] \) has been validated, then \( t \) should be updated with \( t[\text{city}] := s[\text{city}] \). As shown in Fig. 2, the pattern “\( \neq 0800 \)” can be edited via a pop-up frame.

CerFix automatically tests whether the specified eRs make sense w.r.t. master data, i.e., the rules do not contradict each other and will lead to a unique fix for any input tuple. Furthermore, given certain attributes that are validated, it automatically derives what other attributes can be validated by eR and master data, via an inference system.

Data monitor. We shall demonstrate how CerFix interacts with users to find a certain fix for each input tuple.

1. CerFix suggests a set of attributes for the users to validate. The users may either validate these attributes, or opt to assure that some other attributes are correct. The initial suggestions are computed by region finder.
   - As shown in Fig. 3(a), the attributes suggested by CerFix are highlighted in yellow, i.e., area code AC, phone number phn, phone type and product item. The values of the attributes assigned by the users are 201, 075568485, Mobile phone (type 2), and DVD, respectively.

2. If the users opt to validate these attributes, CerFix iteratively applies editing rules and master data to the data, and expands the set of attributes validated.
   - As shown in Fig. 3(b), all attributes that have been validated are now highlighted in green. These include attributes first name FN, last name...
is to compute suggestions. To reduce the cost, CerFix presents the statistics about the attribute FN, namely, the percentage of FN values that were validated by the users and the percentage of values that were automatically fixed by CerFix. Our experimental study indicates that in average, 20% of values are validated by users while CerFix automatically fixes 80% of the data.

**Summary.** The demonstration aims to exhibit the strength of editing rules and important functionalities of CerFix. (1) Editing rules. As opposed to integrity constraints that only detect the presence of errors in the data, editing rules identify what attributes are erroneous and tell us how to correct the errors with master data. (2) Region finder. It tells us to validate an input tuple, what minimal sets of attributes have to be assured correct. (3) Data monitor. It interacts with the users to find certain fixes, while minimizing human efforts by suggesting a minimal number of attributes for the users to validate. (4) Data auditing. It helps the users understand better the quality of input data sets.

**Acknowledgments.** Fan is supported in part by the RSE-NSFC Joint Project Scheme and an IBM scalable data analytics for a smarter planet innovation award. Li is supported in part by NGFR 973 grant 2006CB303000 and NSFC grant 60533110. Shuai is supported in part by NGFR 973 grant 2011CB302602 and NSFC grants 90818028 and 60903149.

### References


