فصلنامه سیستم‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)

Winter 2020, 1 (2), 33-41
DOR: 20.1001.1.27832570.1400.2.2.2.5

Received: 2 Nov 2019
Accepted: 13 Jan 2020

A New Approach for Data Cleaning to Improve Quality of Data Warehouse

Ali Shahnavaz[1], Mehdi Afzali[2*], Shima Rahimzadeh[3]

1. Assistant Professor, Department of Mathematics and Statistics, Zanjan Branch, Islamic Azad University, Zanjan, Iran. shahnavaz_ali2000@yahoo.com
2. Assistant Professor, Department of IT Engineering, Zanjan Branch, Islamic Azad University, Zanjan, Iran.*(Corresponding Author)* afzali@iauz.ac.ir
3.MSc Student, Zanjan University of Medical Sciences, Zanjan, Iran.

Abstract

Introduction: Business intelligence is defined as the process of converting data into information and then knowledge. Business intelligence represents a broad application and technology space for collecting, storing, analyzing and accessing information to improve high-quality business process modeling. Business intelligence consists of several steps. In the first step, data sources are collected. These sources can include data from various types of databases or information from existing software. The information collected during the "ETL" process is loaded into the analytical database or data warehouse.، There have been several works addressing data cleaning, including correcting spelling and phonetic errors, unifying the format of the date field, correcting quantitative data, deleting duplicate records, etc. But based on the research done, nothing has been done to resolve the contradiction in interrelated fields (e.g. geographical locations), and since this field is available in the organization's database, its resolution is a problem for improving the organization's performance. It is important and necessary that it should be addressed.

Method: It is necessary to use the standard data of provinces and cities of Iran, and for this purpose, a database containing the information of the provinces and cities of Iran and the first three digits of the national code for each city was created in the Microsoft Access 2010, and the connection between the related tables was established. The health certificate database of Zanjan University of Medical Sciences students who graduated in 1992 and 1993 was received from the information technology department of Zanjan University of Medical Sciences. These data were based on different semesters for two consecutive years. The proposed program is written in the Visual Studio 2013 environment and has two databases, one is the standard culture of provinces and cities and the three digits of their codes, and the other is a student database for testing and resolving data contradictions.

Finding: Through the implementation of this approach we have been able to detect dirty data and then by using students' national codes, the correction process has been applied to them. Based on the achieved results, the amount of dirty data decreased from 25.79% to 4.97%.

Conclusion: In this article, an attempt was made to provide a new method for data cleaning. In order to achieve the desired goal, especially the data of provinces and cities were cleaned using the prepared standard data. By implementing the proposed method, we were able to identify the contaminated data in this characteristic in the examination of the students' national code and then implement the data correction process on them.

Keywords: Data Preparation, Data warehouse, Data Quality, Data Cleaning, Data Warehouse, Dirty Data, Data Management

فصلنامه سیستم‌های پردازشی و ارتباطی چندرسانه‌ای هوشمند
Intelligent Multimedia Processing and Communication Systems (IMPCS)
انجمن علمی تجارت الکترونیکی ایران
واحد زنجان

ارائه روشی جدید برای پاکسازی داده‌ها جهت بهبود کیفیت انبار داده

تاریخ دریافت: ۱۳۹۹/۰۸/۱۱ دورهٔ اول، زمستان ۱۳۹۹
تاریخ پذیرش: ۱۳۹۹/۱۰/۲۳ شمارهٔ دوم، صص: ۳۳ – ۴۱

علی شهنواز[۱]، مهدی افضلی[۲*]، شیما رحیم‌زاده[۳]

۱. استادیار، گروه ریاضی و آمار، واحد زنجان، دانشگاه آزاد اسلامی، زنجان، ایران. shahnavaz_ali2000@yahoo.com
۲. استادیار، گروه مهندسی فن‌آوری اطلاعات، واحد زنجان، دانشگاه آزاد اسلامی، زنجان، ایران. (نویسنده مسئول) afzali@iauz.ac.ir
۳. کارشناسی ارشد، دانشگاه علوم پزشکی و خدمات درمانی و بهداشتی زنجان، زنجان، ایران.

چکیده: مهمترین مسئله در مدیریت داده‌ها، موضوع کیفیت داده است. کیفیت داده می‌تواند پاکسازی داده‌ها را قبل از بارگذاری به انبار داده‌ها تضمین‌کند. پاکسازی داده فعالیتی است شامل فرآیند تشخیص و اصلاح اشتباهات و تناقضات در انبار داده‌ها. به‌دلیل وجود اطلاعات زیاد در بانک‌های اطلاعاتی مشکلات و تناقضات فراوانی درآن‌ها به‌وجودآمده‌است. هدف اصلی ما ارائه روشی برای رفع تناقضات موجود در بانک‌های اطلاعاتی برای پاک‌سازی داده‌های آلوده می‌باشد. با هدف بهبود کیفیت انبار داده برای تصمیم‌گیری‌های صحیح، روش جدیدی ارائه‌شده‌است و برای آزمایش روش پیشنهادی، از بانک اطلاعاتی شناسنامه سلامت دانشجویان دانشگاه علوم پزشکی زنجان ورودی سال‌های ۹۲ و ۹۳، شامل ۸۴۵ نفر که در حال حاضر همه آن‌ها فارغ‌التحصیل شده‌اند به‌عنوان داده‌های مورد بررسی استفاده‌شده‌است. برنامه پیشنهادی با زبان برنامه‌نویسی سی‌شارپ پیاده‌سازی و اجراشده‌است. برنامه یا اپلیکیشن ما در چهار لایه و به‌صورت ویندوز اپلیکیشن نوشته‌شده‌است. از طریق اجرای روش پیشنهادی توانستیم با بررسی کدملی دانشجویان، داده‌های آلوده در این مشخصه را تشخیص‌داده و سپس فرآیند اصلاح داده را روی آن‌ها اعمال‌نماییم. براساس نتایج به‌دست‌آمده، میزان داده آلوده در انبار داده تولیدشده از ۲۵,۷۹ درصد به ۴,۹۷ درصد کاهش‌یافت.

واژه‌های کلیدی: مدیریت داده، آماده‌سازی، انبار داده، داده‌کاوی، داده‌های‌آلوده، پاک‌سازی.

۱. مقدمه

فن‌آوری‌های نوین با سرعتی سرسام‌آور در حال پیشرفت هستند، آن‌گونه که جوامع به صورت عام و بازار به صورت خاص با شتابی وصف‌ناپذیر به دنبال ترفندهایی می گردند که بقایشان را در این عرصه آشفته و متلاطم تضمین‌کنند. سازمان‌ها باید بپذیرند که فلسفه حیاتشان تغییر کرده‌است و دیگر زنده‌بودن به معنای رسیدن به وضعیت سوددهی مداوم نمی‌تواند باشد و باید به دنبال رقابت و ابزار آن بود. یک سازمان در طول حیاتش، داده ایجادمی‌کند و این حقیقت سازمان‌ها را ملزم به جستجوی ابزارهایی برای تسهیل فرآیند کسب اثربخش داده‌ها، پردازش و تحلیل وسیع آن‌ها کرده‌است تا براساس آن پایه‌ای را برای کشف دانش جدید بنا نهند. بنابراین تسلط بر فن‌آوری‌های جدیدی مانند هوش تجاری[۱] در کسب‌وکارها یک الزام وضرورت اجتناب‌ناپذیر تلقی می‌شود.

هوش تجاری به‌عنوان فرآیند تبدیل داده به اطلاعات و سپس دانش تعریف می‌شود[1]. هوش تجاری نشان‌دهنده فضای گسترده کاربرد و فن‌آوری برای جمع‌آوری، ذخیره‌سازی، تجزیه و تحلیل و دسترسی به اطلاعات برای بهبود مدل‌سازی فرآیند تجاری باکیفیت است.

مراحل هوش تجاری شامل مراحلی است. در مرحله اول، منابع داده جمع‌آوری می‌شوند. این منابع می‌تواند داده‌های انواع پایگاه داده یا اطلاعات نرم‌افزارهای موجود را دربربگیرد. اطلاعات جمع‌آوری شده طی فرآیند "ای تی ال"[۲] در پایگاه داده تحلیلی یا همان انبارداده بارگذاری‌می‌شود.

رالف کیمبال انبار داده‌ها را به‌عنوان "یک کپی از سیستم‌های معاملاتی، به‌ویژه ساختار برای پرس‌وجو و تجزیه و تحلیل" تعریف‌می‌کند. داده در پایگاه داده تحلیلی در بخش‌های مجزایی به نام انبارک[۳] قرارمی‌گیرد. در انبار داده‌ها کل داده‌های جمع‌آوری‌شده وجوددارد و انبارک‌ها شامل تنها یک زیرمجموعه از حجم داده‌های سازمان، مختص به یک گروه خاصی از کاربران، در موضوعی خاص محدود شده‌است[2].

انبارداده‌ها برای اهداف تحلیلی یعنی پردازش تحلیلی آنلاین "اولپ"[۴] استفاده می‌شوند، انبار داده موضوع‌گرا، یکپارچه، غیرفرّار و مغایر زمان جمع‌آوری داده‌ها در حمایت از مدیریت تصمیم‌گیری است[3]. انبار داده‌ها به همراه "ای تی ال" و ابزار گزارش، یک محیط یکپارچه برای پردازش کسب‌وکار ایجاد می‌کند[4]. در مرحله بعد هوش تجاری وارد عمل شده و اطلاعات طبقه‌بندی‌شده را تجزیه و تحلیل می‌کند. درنهایت اطلاعات جهت انتشار به ابزارهای سطح بالا تحویل‌داده‌می‌شود[2].

بیش از ۷۰٪ از پروژه‌های انبار داده‌ها در بخش «ای تی ال» صرف‌می‌شود[5]. امروزه سناریوی ابزار «ای تی ال» قسمت مهمی از مسئولیت نرم‌افزار برای یکپارچه‌سازی اطلاعات ناهمگون از منابع مختلف است [6,7]. انجام فرآیند «ای تی ال» بالقوه پیچیده، سخت و وقت‌گیر است.

شوران (۲۰۱۴) کل مشکلات موجود در داده‌های انبار داده را براساس کیفیت داده دسته‌بندی‌کرده‌است[13]. چودهاری در سال ۲۰۱۴ مشکلات و روش‌های پاکسازی داده‌ها را مطرح کرد.[7] میلانی و دکتر گوپتا (۲۰۱۵) نیز مشکلات کیفیت داده‌ها در مرحله "ای تی ال" را نام‌بردند[9]. این پژوهش‌ها نشان از اهمیت مشکلات موجود است و همچنان تحقیقات روی دسته‌بندی و معرفی این مشکلات ادامه‌دارد.

تاکنون کارهای فراوانی جهت پاک‌سازی داده‌ها صورت‌گرفته‌است، اعم از رفع غلط املایی و آوایی، یکی‌کردن فرمت فیلد تاریخ، اصلاح داده‌های کمّی، حذف رکورد تکراری و غیره. ولی براساس پژوهش انجام‌گرفته، برای رفع تناقض در فیلدهای وابسته به‌هم (مانند: محل‌های جغرافیایی) کاری انجام‌نشده‌است و چون این فیلد در بانک اطلاعاتی سازمان‌ها موجوداست، بنابراین رفع آن برای بهبود عملکرد سازمان‌ها مسئله‌ای مهم و ضروری است که باید به آن پرداخته‌شود. کارهای انجام‌گرفته همیشه از طریق دو فیلد تشخیص و اصلاح خطا انجام‌شده‌است در این پژوهش برآنیم که از طریق سه فیلد وابسته به هم، این کار را انجام‌دهیم. هدف اصلی ارائه روش جدیدی جهت بالابردن کیفیت داده از طریق رفع تناقض در فیلدهای وابسته به هم و پرکردن فیلدهای خالی -در صورت وجود- و حذف رکوردهای تکراری در بانک اطلاعاتی است.

۲. پیشینه پژوهش

وارول و همکاران سال (۲۰۱۰) الگوریتم PNRS[۵] را ارائه‌دادند. این الگوریتم برای تصحیح خطاهای متنی و آوایی، از فرهنگ لغت جهانی استفاده‌کرده و به پاک‌سازی داده می‌پردازد. مشکل روش یادشده این منحصربودن فرهنگ لغت جهانی به زبان انگلیسی است[14].

لی و همکاران (۲۰۱۰) الگوریتم بستار متعدی[۶] را ارائه‌دادند، که در آن برای حذف رکوردهای تکراری و پرکردن مقادیر فیلدهای خالی از کلید برای تطبیق استفاده‌می‌شود اما مشکل این است که تنها یک کلید تطبیق استفاده‌می‌شود و به‌صورت کامل هم خودکار نیست[15].

همّاد و جیهاد (۲۰۱۱) تکنیک‌های پیشرفته[۷] برای پاک‌سازی داده را ارائه‌دادند که جهت تصحیح داده‌های کمّی استفاده‌می‌شود این تکنیک در اصلاح داده‌هایی مثل سن و تاریخ تولد به‌کار می‌رود[16].

پائول و همکاران در سال ۲۰۱۲ روش ترکیبی HADCLEAN را برای پاک‌سازی داده در انبار داده‌ها ارائه‌دادند و در آن از ترکیب اصلاح‌شده نسخه‌های الگوریتم PNRS و الگوریتم بستار متعدی استفاده‌کردند. در روش مذکور به جای فرهنگ لغت جهانی از فرهنگ اصطلاحات مخصوص سازمان استفاده و هجده رکورد بیشتر از الگوریتم اصلی PNRS اصلاح‌شده‌است[17].

پوروال و وورا در سال ۲۰۱۳ دو الگوریتم قوانین انجمنی[۸]، HADCLEAN و روش‌های آن‌ها برای کیفیت داده‌ها را شرح‌داده‌اند. این تحقیق همچنین شامل مقایسه عوامل مختلف و جنبه‌های مشترک در این دو الگوریتم است. الگوریتم قوانین انجمنی با استفاده از محاسبات سنگین ریاضی خطای دورویی از نوع رشته داده (فیلد نام) را می‌یابد[12].

کولکارنی و باکال در سال ۲۰۱۴ هم روش ترکیبی HADCLEAN را برای پاک‌سازی داده در انبار داده‌ها دوباره مطرح‌کرده و شرح‌داده‌اند

[18]. آشوینی و همکاران نیز (۲۰۱۴) از تکنیک‌های ترکیبی برای پاک‌سازی داده‌ها استفاده‌کرده و با ترکیب سه الگوریتم PNRS، تکنیک‌های پیشرفته، و بستار متعدی توانسته‌اند چندین فیلد را با هم پاک‌سازی‌کنند [19].

باتاچارجی و همکاران نیز سال ۲۰۱۴ روی پاک‌سازی بانک اطلاعاتی براساس «ای تی ال» تحقیق کردند. آن‌ها با اعمال الگوریتم‌های نسبیت موفق، حذف رکوردهای تکراری، ادغام جداول به‌ترتیب باعث کاهش خطا در بانک اطلاعاتی شدند که درنهایت بعد از اعمال الگوریتم، ۳٪ از کل خطا، باقی‌ماند[8].

تانجا و همکاران در سال ۲۰۱۴ الگوریتم نوین DFT[۹] (تبدیل فیلد تاریخ) را برای پاکسازی داده‌ها ارائه کردند. این الگوریتم که با استفاده از جاوا پیاده‌سازی شده‌است و مشکلات ذکرشده در فیلد تاریخ را اصلاح می‌کند[10].

دکتر کالیا و دوی (۲۰۱۵) به معرفی و مقایسه ابزارهای پاکسازی داده اعم از: RAPIDMINOR-WINPURE CLEAN & MATCH - MS EXCEL DATA CLEANER پرداختند[11].

۳. روش‌شناسی پژوهش

استفاده از فرهنگ استاندارد داده‌های استان‌ها و شهرستان‌های ایران ضروری است و برای این کار ابتدا یک بانک اطلاعاتی شامل اطلاعات استان‌ها و شهرستان‌های ایران و سه رقم اول کدملی مخصوص هر شهرستان، در محیط اکسس ۲۰۱۰ ساخته‌شد و ارتباط میان جداول مربوط برقرارشد. سپس بانک اطلاعاتی حاصل از محیط اکسس به اس کیو ال سرور ۲۰۱۴ انتقال‌داده‌شد (این کار به‌صورت پویا[۱۰] انجام‌شده و قابلیت اضافه‌کردن اطلاعات جدید و یا اصلاح اطلاعات موجود و همچنین امکان حذف از بانک اطلاعاتی در مواقع خاص در آن وجوددارد).

بانک اطلاعاتی شناسنامه سلامت دانشجویان دانشگاه علوم پزشکی زنجان ورودی سال‌های ۹۲ و ۹۳ که فارغ‌التحصیل شده‌اند به‌عنوان داده‌های مورد بررسی از بخش فن‌آوری اطلاعات دانشگاه علوم پزشکی زنجان دریافت‌شد. این داده‌ها براساس ترم‌های مختلف برای دو سال متوالی بود که پس از دریافت، در محیط اکسل ۲۰۱۰ یکپارچه گردید و مجموعاً شامل ۸۴۵ رکورد شد. سپس بانک اطلاعاتی دانشجویان از محیط اکسل به اس کیو ال سرور ۲۰۱۴ انتقال‌داده‌شد. لازم به ذکر است که برنامه پیشنهادی با زبان برنامه‌نویسی سی‌شارپ پیاده‌سازی و اجرا شده‌است (دلیل استفاده از سی‌شارپ این است که دارای زبان مفهومی‌تری برای کاربر است و در آن از دات نت فریم ورک[۱۱] که زیربنای زبان‌های برنامه‌نویسی است، استفاده‌شده‌است (دات نت فریم ورک تمام لایه‌های پیاده‌سازی نرم‌افزار را از سطح سیستم‌عامل به بالا، پوشش‌می‌دهد و فریم ورک مجموعه‌ای از فایل‌های مورد نیاز سیستم‌عامل است که برای اجرای برنامه نوشته‌شده تحت دات نت ضروری است). برنامه پیشنهادی در محیط ویژوال استودیو[۱۲] ۲۰۱۳ نوشته‌شده و دارای دو بانک اطلاعاتی یکی فرهنگ استاندارد استان‌ها و شهرها و سه رقم کد مربوط به آن‌ها و دیگری بانک اطلاعاتی دانشجویی برای انجام آزمایش و رفع تناقضات داده‌هاست، هر دو بانک اطلاعاتی در اس کیو ال سرور ۲۰۱۴ هستند (دلیل استفاده از اس کیو ال سرور هم این است که این نسخه تقریباً با بانک اطلاعاتی اوراکل برابری می‌کند و در میلیاردها رکورد، بسیار سریع پاسخ‌می‌دهد). سپس ارتباط برنامه با این دو بانک اطلاعاتی از طریق دیتاست انجام‌می‌شود، یعنی برای ارتباط از کلاس دیتاست استفاده‌شده‌است. برنامه یا اپلیکیشن ما در چهار لایه و به‌صورت ویندوز اپلیکیشن[۱۳] نوشته‌شده‌است:

۱) لایه ارتباط با بانک که جهت ارتباط اپلیکیشن با بانک اطلاعاتی استفاده‌می‌شود.

۲) لایه دستورات جستجو و غیره، اسکریپت‌های مختلف جستجو در این لایه نوشته‌می‌شود. اسکریپت‌ها می‌توانند مستیقماً در محیط برنامه نوشته یا به‌صورت store procedure از بانک اطلاعاتی فراخوانی شوند.

۳) لایه ارتباط لایه اول و دوم؛

۴) لایه دستورات زبان برنامه‌نویسی، که بخش اصلی برنامه بوده و طراحی اینترفیس[۱۴] برنامه هم در این لایه انجام‌می‌گیرد.

در برنامه پیشنهادی از کامپوننت Janus Winforms Controls v4 جهت زیباسازی خروجی برنامه استفاده‌شده‌است.

۱،۳. شرح مراحل برنامه

ابتدا کلیه اطلاعات فرهنگ استاندارد استان‌ها و شهرستان‌ها و سه رقم کدملی هر شهرستان خوانده و در یک دیتاست ذخیره‌می‌شود.

سپس کلیه اطلاعات بانک اطلاعاتی دانشجویان خوانده‌می‌شود و در دیتاست دیگری ذخیره‌می‌شود که حاصل کار اطلاعات زیر است:

استان، شهرستان و کدملی.

در مرحله بعد، برنامه می‌سنجد که آیا فیلد کد ملی دارای مقدار می‌باشد یا مقدار فیلد کد ملی null و یا blank است؟ اگر خیر آن سطر از جدول دانشجویان را در جدول داده‌های فاقد کدملی ذخیره می‌کند و در خروجی نمایش می‌دهد. اگر بله به مرحله بعدی می‌رود و ده‌تایی بودن تعداد کاراکترهای فیلد کدملی را می‌سنجد. اگر بله بود، ایده به‌کاررفته در این مرحله استفاده از جدولی به نام temp است. جدول موقتی که در اس کیو ال سرور ساخته‌شده داده‌های آلوده کشف‌شده در این بخش را از جدول اصلی دانشجویان می‌خواند و آن‌ها را در جدول temp کپی‌می‌کند و در خروجی نمایش‌می‌دهد. این جدول موقت در هر مرحله و هر بار که فراخوانی می‌شود، ابتدا همه داده‌های خود را پاک‌می‌کند و از اول آن را بسته به دستور خواسته‌شده از داده‌های آلوده پرمی‌کند و نمایش‌می‌دهد. حال اگر تعداد کاراکتر کدملی مطابق انتظار ده بود و درست بود به مرحله بعدی می‌رود و می‌سنجد که فرمت فیلد کدملی درست باشد. اگر چنین بود، باز هم از جدول temp استفاده‌می‌کند و اطلاعات قبلی را پاک کرده و داده‌های آلوده جدید را در آن می‌ریزد و نمایش‌می‌دهد.

اما خطای کدملی چگونه تشخیص‌داده‌می‌شود؟ کد ملی شماره‌ای است ده رقمی که از سمت چپ سه رقم کد شهرستان محل صدور

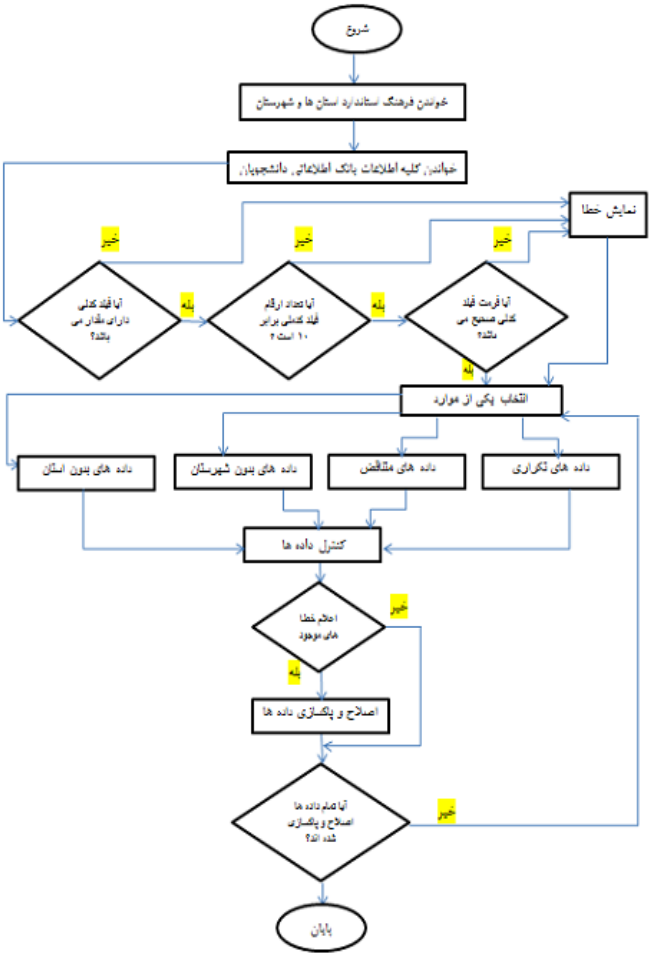
شناسنامه، شش رقم بعدی کد منحصربه‌فرد برای فرد دارنده شناسنامه در شهرستان محل صدور و رقم آخر هم یک رقم کنترل است که از روی نه رقم سمت چپ به‌دست‌می‌آید. برای بررسی کنترل کد کافی است مجدداً از روی نه رقم سمت چپ رقم کنترل را محاسبه‌کنیم. برای محاسبه رقم کنترل از روی سایر ارقام:

- هر رقم را در موقعیت آن ضرب‌کرده و حاصل را با هم جمع‌می‌کنیم.
- مجموع به‌دست‌آمده از مرحله یک را بر عدد یازده تقسیم‌می‌کنیم.
- اگر باقیمانده کمتر از عدد دو باشد، رقم کنترل باید برابر باقیمانده باشد در غیر این صورت، رقم کنترل باید برابر یازده منهای باقیمانده باشد.

این فرمول کلی در برنامه برای تک تک کدملی‌ها محاسبه‌می‌گردد و به‌وسیله آن داده‌های آلوده این بخش مشخص می‌گردد.

در مرحله بعد داده‌های تکراری، داده‌های بدون استان، داده‌های بدون شهرستان و داده‌هایی که استان و شهرستان آن‌ها با هم تناقض‌دارند، کنترل و محاسبه‌می‌شوند.

ایده به‌کاررفته برای کشف تکرار چنین است: کدملی‌هایی که کاراکتر به کاراکتر عین هم باشند را می‌یابد و باز هم با استفاده از جدول temp آن‌ها را نمایش‌می‌دهد. برای سه قسمت بعدی در قسمت view اس کیو ال سرور یک جدول جدید از ترکیب جداول فرهنگ استاندارد استان‌ها و شهرستان‌ها ساخته‌ایم که نام استان‌ها و شهرستان‌ها و سه رقم کد مربوط به آن شهرستان‌هاست و داده‌های آلوده استان و شهرستان با توجه به این جدول ترکیبی ساخته‌شده استخراج‌می‌شود، جدول موقت temp با این خطاها پرشده و در خروجی نشان‌داده‌می‌شود.



شکل ۱: فلوچارت برنامه پیشنهادی

پس از کشف این داده‌های آلوده وارد مرحله بعد می‌شویم و اصلاح و پاک‌سازی داده‌ها آغازمی‌شود. تا قبل از این مرحله خطاها فقط تشخیص‌داده‌می‌شدند، پس از این مرحله وارد فاز پاک‌سازی داده‌ها می‌شویم، اصلاح داده‌های آلوده با استفاده از فرهنگ استاندارد استان‌ها و شهرستان‌ها آغازمی‌شود و داده‌های تکراری نیز از بین می‌روند و فقط یک نسخه از آن‌ها باقی‌می‌ماند. در این برنامه همچنین اگر داده‌ها دارای غلط املایی باشند به‌عنوان داده آلوده شناسایی و با توجه به فرهنگ استاندارد تهیه‌شده، اصلاح‌می‌شوند. ازآنجاکه فرهنگ استاندارد طراحی و ساخت خود ماست، عاری از هرگونه غلط املایی است و می‌تواند این نوع از خطاها را هم تشخیص‌داده و اصلاح‌کند. لازم به ذکر است که در هر مرحله خطاها با ذکر تعداد در خروجی نمایش داده‌می‌شود.

برنامه پس از اتمام مرحله پاک‌سازی داده، در مرحله بعد می‌سنجد که همه داده‌ها از جدول خوانده‌شده‌باشند که در این صورت، کار به‌اتمام‌می‌رسد و داده‌های اصلاح‌شده نمایش‌داده‌می‌شوند.

۲،۳. شرح روند اجرایی و خروجی‌های برنامه

برنامه پیشنهادی یک صفحه اصلی و سه منو دارد: ۱) بانک استان‌ها ۲) بانک افراد ۳) تحلیل داده‌ها.

۱) بانک استان‌ها: در این بخش، بانک اطلاعاتی فرهنگ استاندارد استان‌ها و شهرها به‌نمایش درآمده‌است که در فرم ابتدایی آن نام تمامی سی و دو استان ایران ثبت‌شده‌است و اگر وارد ثبت شهرستان هر کدام از استان‌ها شویم، شهرستان‌های آن استان را در فرم بعدی به‌صورت کامل نمایش‌می‌دهد. تعداد کل شهرستان‌های ثبت‌شده ۵۱۴ شهرستان است. اگر وارد ثبت کدهای شهرستان هر استان شویم، سه رقم اول کدملی مخصوص آن شهرستان را در فرم بعدی می‌بینیم. هر شهرستان ممکن‌است دارای یک یا چندین کد شهرستان باشد، به‌عنوان مثال شهرستان اسلامشهر دارای یک کد شهرستان است ولی تهران دارای پانزده کد شهرستان می‌باشد. که ایده به‌کار رفته در کدنویسی برنامه برای جداکردن کدهای شهرستان استفاده از کاراکتر * (ستاره) است. تعداد کل کد شهرستان‌های ثبت‌شده ۶۵۵ مورد است. لازم به ذکر است که فرهنگ استان و شهرستان تهیه‌شده براساس آخرین تقسیم‌بندی جغرافیایی کشور ایران است.

در این فرم‌ها هر کدام از استان‌ها و شهرستان‌ها و کدهای آن‌ها دارای قسمت‌های حذف و ویرایش و اضافه‌کردن مورد است یعنی به‌صورت پویا، اگر زمانی تقسیم‌بندی کشور دچار هرگونه تغییری شود، امکان اعمال آن در فرهنگ استاندارد مذکور تعبیه‌شده‌باشد و نیز اگر به هر دلیلی بانک اطلاعاتی دچار اختلال‌شد، اضافه، ویرایش و حذف رکوردهای آن امکان‌پذیر باشد.

شکل ۲: خروجی برنامه در بخش بانک استان‌ها

۲) بانک افراد: قسمت بانک افراد دارای دو بخش انتخابی برای استان و شهرستان است. ایده به‌کاررفته در این بخش چنین است که اگر زمانی استانی فاقد شهرستان باشد، برای اینکه پیغام خطا ندهد و از اجرای برنامه خارج‌نشود، از حلقهٔ try، catch استفاده‌شده‌است. وقتی در بخش‌های انتخابی استان و شهرستانی را انتخاب‌می‌کنیم، تمام داده‌هایی را که استان و شهرستان آن‌ها درست است، طبق انتخاب ما از بانک دانشجویان می‌خواند و در جدولی با مشخص‌کردن تعداد رکوردهای موجود نمایش‌می‌دهد. این بخش برنامه هم دارای گزینه ویرایش و حذف می‌باشد.

شکل ۳: خروجی برنامه در بخش بانک افراد

۳) تحلیل داده‌ها: این قسمت دارای هشت قسمت انتخابی می‌باشد شامل نمایش کلیه داده‌ها، نمایش داده‌های تکراری، داده‌های بدون کدملی، داده‌های با فرمت نادرست کدملی، داده‌های با تعداد ارقام نادرست کدملی، داده‌های بدون استان، داده‌های بدون شهرستان و داده‌های دارای استان و شهرستان متناقض، که براساس انتخاب ما در خروجی داده‌های مربوطه را به‌همراه تعدادشان نمایش‌می‌دهد. پس از کشف داده‌های آلوده از طریق دکمه اصلاح، می‌توان خطاها را اصلاح‌نمود.

شکل ۴: خروجی برنامه در قسمت تحلیل داده‌ها

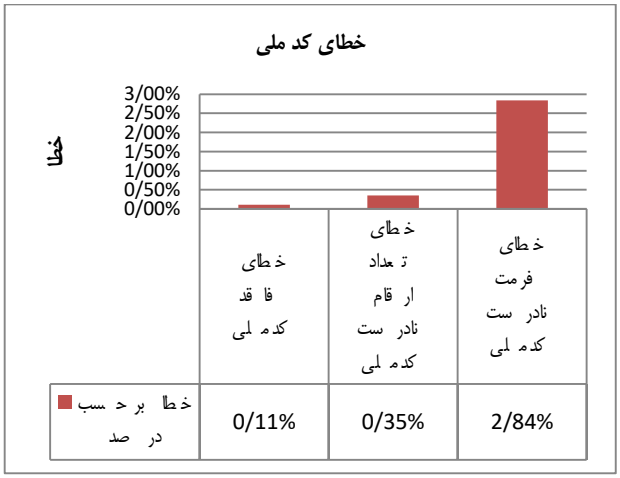
۴. پیاده‌سازی روش پیشنهادی

در این مرحله پیاده‌سازی برنامه آغاز می‌شود و همان‌طور که قبلاً گفته‌شد روش پیشنهادی را روی جامعه آماری یعنی بانک اطلاعاتی مربوط به شناسنامه سلامت دانشجویان، اعمال‌می‌کنیم. داده‌های آلوده به‌طور کامل شناسایی‌می‌شوند و تعداد داده‌های آلوده مربوط به هر بخش در برنامه مشخص‌می‌شود.

خطاهای فاز اول پیاده‌سازی برنامه شامل این سه نوع خطای زیر است:

۱،۴. خطاهای مرتبط با کد ملی

بانک اطلاعاتی دانشجویان تنها دارای یک رکورد فاقد کد ملی می‌باشد، بر این اساس ۰،۱۱ درصد از خطا را به خود اختصاص داده‌است. تعداد سه رکورد وجود دارد که دارای تعداد ارقام نادرست فیلد کد ملی می‌باشند، لذا ۰،۳۵ درصد از خطا مربوط به این نوع از داده‌های آلوده می‌باشد. تعداد ۲۴ رکورد دارای فرمت نادرست کد ملی برطبق فرمول‌بندی تشخیص صحت کد ملی هستند که ۲،۴۸ درصد از خطا را به خود اختصاص‌داده‌اند.

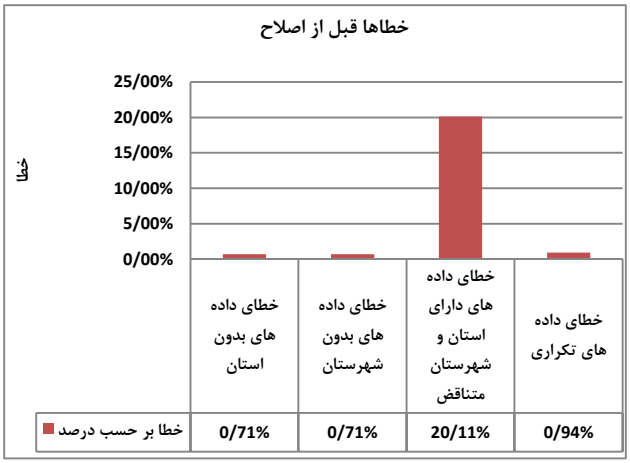


شکل ۵: نمودار خطاهای مربوط به کد ملی

۲،۴. خطاهای مربوط به داده‌های استان‌ها و شهرستان‌ها

داده‌هایی که فاقد استان می‌باشند، شش مورد کشف‌شده‌اند و ۰،۷۱ درصد از خطا مربوط به این نوع از داده‌هااست. داده‌هایی که فاقد شهرستان می‌باشند، هم شش مورد است که باز هم ۰،۷۱ درصد از خطا را به خود اختصاص‌داده‌اند. داده‌هایی که استان و شهرستان آن‌ها با هم در تناقض‌اند، تعداد ۱۷۰ داده است و ۲۰،۱۱ درصد خطا مربوط به این نوع از داده‌ها می‌باشد. داده‌های تکراری هشت مورد است و ۰،۹۴ درصد از خطا را به خود اختصاص‌داده‌است.

به این ترتیب فاز دوم تشخیص خطا هم به پایان می‌رسد.



شکل شماره ۶: نمودار خطای کشف شده در برنامه قبل از اصلاح

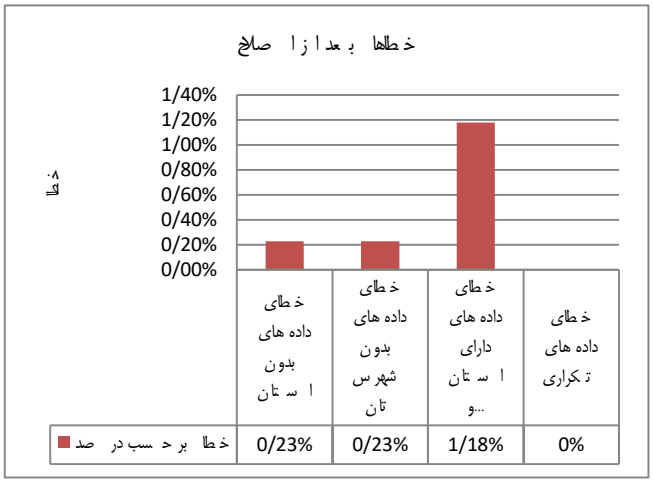
۳،۴. مرحله پاکسازی داده‌ها

پس از اتمام فاز تشخیص خطا در برنامه نوبت به پاکسازی این داده‌های آلوده می‌رسد. کل فرآیند پاک‌سازی داده‌ها در برنامه پیشنهادی ده ثانیه زمان می‌برد. و تمامی داده‌های آلوده شناسایی شده در فاز دوم تشخیص، به‌طور کامل اصلاح‌می‌شوند و پس از آن داده‌های اصلاح‌شده را به ما نمایش‌می‌دهد.

چون روش پیشنهادی ما شامل اصلاح خطاهای مربوط به صحت کدملی نیست، به‌همین دلیل پس از مرحله اصلاح تمام خطاها به صفر درصد نمی‌رسند. و ما به طور دقیق تاثیر این خطاها بر درصد اصلاح را محاسبه‌نموده‌ایم، که در داده‌های بدون استان و شهرستان دو رکورد دارای خطای صحت کدملی هستند، پس خطا از ۰،۷۱ درصد به ۰،۳۵ درصد کاهش‌می‌یابد.

در داده‌هایی که دارای استان و شهرستان متناقض هستند، تعداد ده خطای صحت کدملی داریم که با این حساب خطا از ۲۰،۱۱ درصد به ۱،۱۸ درصد می‌رسد.

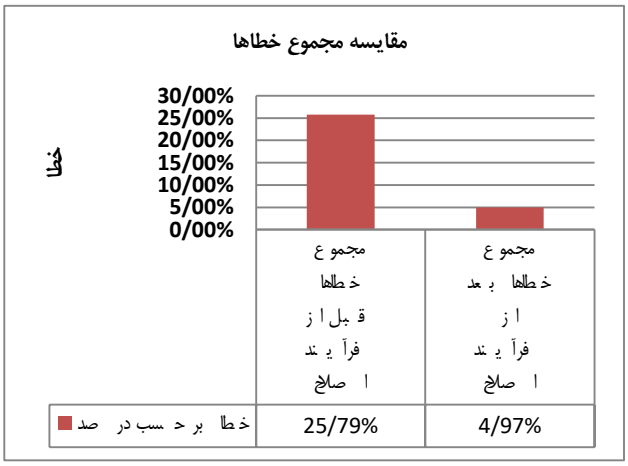
در داده‌های تکراری چون خطای صحت کدملی وجود ندارد، خطا از ۰،۹۴ درصد به صفر درصد می‌رسد.



شکل ۷: نمودار خطای باقی‌مانده پس از اصلاح و پاکسازی داده‌ها

۴.۴. مجموع خطاهای کشف‌شده قبل و بعد از مرحله پاکسازی داده‌ها

کل داده‌های جامعه آماری ما دارای ۸۴۵ رکورد است و مجموع خطاهای محاسبه‌شده در هر فاز ٪۲۵,۷۹ می‌باشد و پس از مرحله پاک‌سازی داده‌ها، مجموع خطاها را به ٪۴,۹۷ رساندیم.



شکل ۸: نمودار مقایسه مجموع خطاها قبل و بعد از فرآیند اصلاح

مقایسه مجموع خطا قبل و بعد از فرآیند اصلاح و پاک‌سازی داده‌ها نشان‌می‌دهد که چقدر توانسته‌ایم میزان مجموع داده‌های آلوده را کاهش‌دهیم.

جدول ۱: خطای داده‌ها در نگاه کلی

خطا	قبل از اصلاح	بعد از اصلاح	میزان اصلاح انجام‌شده
داده فاقد کد ملی	٪۰,۱۱	٪۰,۱۱	٪۰
داده دارای کد ملی با تعداد ارقام نادرست	٪۰,۳۵	٪۰,۳۵	٪۰
داده دارای کد ملی با فرمت نادرست	٪۲,۸۴	٪۲,۸۴	٪۰
داده‌های فاقد استان	٪۰,۷۱	٪۰,۲۳	٪۰,۴۸
داده‌های فاقد شهرستان	٪۰,۷۱	٪۰,۲۳	٪۰,۴۸
داده‌های دارای استان و شهرستان متناقض	٪۲۰,۱۱	٪۱,۱۸	٪۱۸,۹۳
داده‌های تکراری	٪۰,۹۴	٪۰	٪۰,۹۴
مجموع داده‌ها	**٪۲۵,۷۹**	**٪۴,۹۷**	**٪۲۰,۸۲**

تمام خطاهای کشف‌شده قبل و بعد از فرآیند اصلاح و نیز میزان اصلاحی که روش پیشنهادی توانسته انجام‌دهد در جدول زیر آورده‌شده‌اند. نمونه‌ای از داده‌های آلوده کشف‌شده قبل و بعد از فرآیند اصلاح و پاک‌سازی داده‌ها در شکل‌های ۹ و ۱۰ آورده‌شده‌است.

شکل ۹: خروجی برنامه قبل از پاکسازی داده‌های آلوده

شکل ۱۰: خروجی برنامه بعد از پاکسازی داده‌های آلوده

جهت راستی‌آزمایی روش پیشنهادی، کشف و پاکسازی داده‌های آلوده را به‌صورت دستی نیز انجام‌دادیم که البته بسیار زمان‌بر بود. همچنین با نهایت دقتی که برای انجام فرآیند به‌صورت دستی داشتیم، در انتها مشاهده‌شد که باز هم نتوانسته‌ایم برخی از داده‌های آلوده را کشف و پاکسازی نماییم، این امر نشان‌دهنده آن است که وجود روش‌هایی این‌چنینی برای تضمین کیفیت در انبار داده و تصمیم-گیری‌های بعدی وابسته به این داده‌ها بسیار ضروری می‌باشد.

۵. نتیجه‌گیری

در این مقاله تلاش بر این بود تا روش جدیدی برای پاکسازی داده‌ها ارائه‌شود. در راستای دست‌یابی به هدف موردنظر، به‌طورخاص پاکسازی روی داده‌های استان‌ها و شهرستان‌های با به‌کارگیری فرهنگ استاندارد تهیه شده انجام شد. برای آنکه نتایج واقعی باشد از اطلاعات دانشجویان دانشگاه علوم پزشکی زنجان استفاده و فرآیند پاکسازی روی آن اعمال گردید. در این پژوهش سه فیلد وابسته به‌هم را در نظرگرفته و همچنین صحت فیلد کد ملی را بررسی‌کرده‌ایم و در پایان خطای حاصل از آن را محاسبه‌کرده‌ایم. روش ارائه‌شده از دقت مناسبی برخوردار است، و به میزان قابل‌توجهی میزان داده های آلوده را کاهش‌داده‌است. هدف اصلی بالابردن کیفیت داده از طریق رفع تناقض در فیلدهای وابسته به هم، پرکردن فیلدهای خالی احتمالی و حذف رکوردهای تکراری در بانک اطلاعاتی بود و روش پیشنهادی توانست ما را به هدف اصلی برساند و سطح کیفیت مطلوبی که در ابتدای کار از ارائه این روش انتظار داشتیم کاملاً تضمین‌کند.

با اجرای روش پیشنهادی توانستیم در بررسی کد ملی دانشجویان، داده‌های آلوده در این مشخصه را تشخیص‌داده و سپس فرآیند اصلاح داده را روی آن‌ها پیاده‌کنیم. با اعمال روش پیشنهادی و براساس نتایج به‌دست‌آمده، میزان داده آلوده در انبار داده تولیدشده از ۲۵,۷۹٪ به ۴,۹۷٪ کاهش‌یافت.

مراجع

[1] Golfarelli, Matteo "New Trends in Business Intelligence". Proceedings of the 28th International Convention MIPRO (BIS&DE&ISS). MIPRO (May 30-June /2005), Opatija, Croatia. PP: (15-20)

[2] NEDELCU, Bogdan. "Business Intelligence Systems". Database Systems Journal. Vol IV. no.4(2013). PP: (12-20)

[3] Inmon, William H. "Building the Data Warehouse". Wiley Publishing. 4th Edition. (2005). P:32

[4] Ghosh, Ranak; Halder, Sujay; Sen, Soumya. "An Integrated Approach to Deploy Data Warehouse in Business Intelligence Environment". IEEE Third International Conference. (7-8 Feb- 2015), Hooghly). PP: (1 – 4)

[5] Talebzadeh, Hossein. A Service-Based Framework for ETL Process Based on Metadata. Journal of Basic and Applied Scientific Research. (2/1/ 2012). PP:(54-59)

[6] Gill, Rupali; Singh, Jaiteg. "A Review of Contemporary Data Quality Issues in Data Warehouse ETL Environment". Journal on Today's Ideas Tomorrow's Technologies. Vol. 2. No. 2. (19 December 2014). PP: (153_160)

[7] Choudhary, Nidhi. "A Study over Problems and Approaches of Data Cleansing/Cleaning", International Journal of Advanced Research in Computer Science and Software Engineering. Vol. 4, Issue 2. (February 2014). PP: (774 _779)

[8] Bhattacharjee, Arup Kumar ; Chatterjee, Partha ; Prasad Shaw, Mukesh ; Chakraborty, Manomoy. "ETL based Cleaning on Database", International Journal of Computer Applications.Vol.105, No. 8. (November 2014). PP: (34– 40)

[9] Miglani, Sakshi; Dr. Gupta, Neha. "An Overview On Evocations Of Data Quality at ETL Stage", International Journal of Advanced Technology in Engineering and Science. Vol. No. 03. Special Issue No. 01. (March 2015). PP: (1429- 1436)

[10] Taneja, Shweta; Ashri, Ishita; Gupta, Shipra; Sharma, Mehak. DFT: "A Novel Algorithm for Data Cleansing", International Journal of Computer Science and Information Technologies. Vol..5 (2). (2014). PP: (2297- 2301)

[11] Devi, Sapna; Dr. Kalia, Arvind. "Study of Data Cleaning & Comparison of Data Cleaning Tools", International Journal of Computer Science and Mobile Computing. Vol. 4. Issue. 3. (March 2015). PP: (360 – 370)

[12] Porwal, Sonal; Vora, Deepali. "A Comparative Analysis of Data Cleaning Approaches to Dirty Data". International Journal of Computer Applications. Vol. 62, No.17. (January 2013). PP: (30- 34)

[13] Sheoran, Jyoti. "Issues of Data Quality in Data Warehouses", International Conference on Advances in Computer Engineering & Applications (ICACEA-2014 at IMSEC.GZB). PP: (6 – 8)

[14] Varol, Cihan; Bayrak, Coskun; Wagner, Rick; Goff, Dana. "Application of the Near Miss Strategy and Edit Distance to handle Dirty Data", International Series in Operations Research & Management Science. Springer US. Vol. 32, (2010). PP: (91 -101)

[15] Ning Li, Wing; Bheemavaram, Roopa; Zhang, Johnson. "Transitive Closure of Data Records", Application and Computation. International Series in Operations Research & Management Science. Springer US. vol. 132. (2010). PP: (39-75)

[16] Dr. M. Hamad, Mortadha; Jihad, Alaa Abdulkhar. "An Enhanced Technique to clean Data in the Data Warehouse". Developments in E-system Engineering (DeSE). IEEE International Conference on (6-8 Dec. 2011). Dubai. PP: (306-311)

[17] Paul, Arindam; Ganesan, Varuni; Challa, Jagat Sesh; Sharma, Yashvardhan. " HADCLEAN: A Hybrid Approach to Data Cleaning in Data Warehouses", Information Retrieval & Knowledge Management (CAMP). IEEE International Conference (13-15 March 2012). Kuala Lumpur. PP: (136- 142)

[18] Kulkarni, Prerna S; Bakal, J.W. "Hybrid Approaches for Data Cleaning in Data Warehouse". International Journal of Computer Applications. Vol 88 , No.18. (February 2014). PP: (7-10)

[19] M. Save, Ashwini; Kolkur, Seema." Hybrid Technique for Data Cleaning". National Conference on Role of Engineers (in Nation Building. 2014. NCRENB-14). PP: (4-8)

پی‌نوشت

1. Business Intelligence
2. ETL (Extract, Transform, Load)
3. Data Mart
4. OLAP (On-Line Analytical Processing)
5. Personal Name Recognizing Strategy
6. Transitive Closure Algorithm
7. Enhanced Technique
8. Alliance rules algorithms
9. Date Field Transformation Algorithm
10. dynamic
11. .Net Framework
12. Visual Studio
13. Windows Application
14. interface