بسم الله الرحمن الرحيم

∞∞∞

تم رفع هذه الرسالة بواسطة / مني مغربي أحمد

بقسم التوثيق الإلكتروني بمركز الشبكات وتكنولوجيا المعلومات دون أدنى مسئولية عن محتوى هذه الرسالة.

ملاحظات: لا يوجد

Information System Department
Faculty of Computer and Information Sciences
Ain Shams University

Ontology based System for Converting Semi Structured Data into Relational Data

A thesis submitted as partial fulfillment of the requirements for the degree of Master in Computer and Information Sciences

By

Arwa Abd Elrahman Abd Elslam Awad

Information Systems Department
Faculty of Computer and Information Sciences
Ain Shams University

Under Supervision of

Prof. Dr. Mohamed Ismail Roushdy
Professor of Computer Science
Faculty of Computer and Information Sciences, Ain Shams University
Dean of Faculty of Computer and Information Technology, Future University in Egypt

Prof. Dr. Ibrahim Fathy Moawad
Professor of Information Systems
Faculty of Computer and Information Sciences, Ain Shams University
Dean of the Faculty of Computer Science and Engineering, New Mansoura University

Prof. Dr. Rania Abd Elrahman Elgohary
Professor of Information Systems
Faculty of Computer and Information Sciences, Ain Shams University
Dean of Faculty of Information Technology, Misr University for Science and Technology

May – 2022
Cairo

Acknowledgment

I would like to thank my supervisors, Prof. Dr. Mohamed Roushdy, Prof. Dr. Ibrahim Moawad, and Prof. Dr. Rania ElGohary for all their continued support and encouragement for me on the thesis during the whole years, also, I would like to thank them for their wise suggestions as well as giving me this chance to take part in the project of the ontology-based system for converting semi-structured data into relational data.

At the same time, I want to thank my testers, Prof. Dr. Abu Ela Hassanein, Prof. Dr. Mohamed Hashem, and Prof. Dr. Mohamed Roushdy for their evaluation and their valuable advice for further thesis improvement.

I would like to thank my parents, for their support and encouragement when I suffered the difficulties of computation during my master time.

At last, the thanks go to my husband, Eng. Ahmed Mohamed Fathy, for his patience during my master time and his continuous support for me all the time.

Abstract

Spreadsheets are contained critical information on various topics and are most broadly utilized in numerous fields. There are a huge amount of spreadsheets clients around the world as it considered the standard documentation format when dealing with data in a tabular format as a result of their convenience, support for diagrams, graphs and gives their users an enormous level of opportunity in encoding their data as it is simple to utilize.

A spreadsheet is designed to work similarly to a database as it has a cell-like structure with a cell being a member of a horizontal row and vertical column. While compared to a database, spreadsheets lack many features that make them less appealing for use in data storage and processing. Spreadsheets suffer from low quality because of duplication or data redundancy may be found, where multiple copies of the same data exist in the same spreadsheet document. In addition, spreadsheets do not have the capacity to provide multiuser access like a database and it also has limited storage capabilities.

Spreadsheet tables with semi-structured form are a type of nearly relational data that shares the important qualities of relational data but does not present itself in a relational format. It often conveys highly valuable information and is widely used in many different areas. If we can convert such data into the relational form, many existing tools can be leveraged for a variety of interesting applications, such as data analysis with relational query systems and data integration applications.

In addition, the methodology wherein information is stored in Excel spreadsheets is not the best way to deal with sorting out and getting to it. As the result of utilizing a spreadsheet is in developing exponentially in the most recent years and the increments in volume and unpredictability of this information have prompted expanded prerequisites to save. By converting a spreadsheet document into a relational format or into a database table, the user can benefit from the advantages of a database on their existing data.

The thesis aims to automate the conversion of the spreadsheet tables from semi-structured format with low quality into high-quality relation form based on ontology technique. We have developed a system that automates converts the semi-structured data (tables) in spreadsheets into relational data without user previous experience in any programming language and converts from Low-Data Quality (LDQ) to High-Data Quality (HDQ). The proposed approach used novel algorithms based on a clustering approach, cell classification strategy, and heuristic rules for table detection and extraction from a spreadsheet. Finally, the tests show that the methodology builds the adaptability of information integration systems. In addition, experiments result achieved high accuracy in extracting relational data from spreadsheets with a percentage of 97.5% and 82.4% in simple and hierarchal data respectively, besides a 100% percent of successfully extracted duplicated records from spreadsheets if found.

List of Publications

1. "Heuristic Algorithm for Automatic Extraction Relational Data from Spreadsheet Hierarchical Tables", Arwa Awad, Rania Elgohary, Ibrahim Moawad and Mohamed Roushdy, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 12, No. 10, 2021, 743-748.

2. "Metadata Extraction for Low-Quality Semi-structured Spreadsheets", Awad, A., Elgohary, R., Moawad, I., Roushdy, M., Advances in Intelligent Systems and Computing, Springer, Vol. 1153, pp. 448-457, Proceeding of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)

3. "An Interactive Tool For Extracting Low-Quality Spreadsheet Tables And Converting Into Relational Database", Arwa Awad, Rania ElGohary, Ibrahim Moawad, Mohamed Roushdy, International Journal of Intelligent Computing and Information Sciences, Vol. 21, No.1, 2021, pp. 1-18

Table of Contents

List of Figures

List of Tables

List of Abbreviations

- **LDQ:** Low Data Quality
- **HDQ:** High Data Quality
- **DQM:** Data Quality Management
- **PK:** Primary Key
- **FK:** Foreign Key
- **XML:** eXtensible Markup Language
- **HTML:** Hypertext Markup Language

List of Algorithms

Chapter1. Introduction

Chapter 1. Introduction

1.1 Overview

Semi-structured data (such as spreadsheets) refers to a type of nearly relational data which shares important qualities of relational data but does not present itself in a relational format. Semi-structured data often contains highly valuable information and is widely used in many different areas. If we can convert such data into a relational format, many existing tools can be leveraged for a variety of interesting applications, such as data analysis with relational query systems, reusing, and sharing for data integration applications.

The features of the semi-structured data can use tags, markers, or standard structures to encode the semantic units. In recent years, a variety of tools, such as SQLServer and MySQL have been developed to manage data in a relational format. Some other tools can query the relational data or integrate with other data sources. These all tools require the data stored in a relational format. Unfortunately, a massive amount of data stored in a spreadsheet is often not in a relational format, which makes it difficult to use most of the existing tools to manage the data.

A spreadsheet is an interacting application tool for organization charts, storage, and analysis of data that is diverse and widely used. Microsoft Excel is the most available, unofficial estimations estimating the number of users to 1.2 billion [1]. Thus, spreadsheets are not only powerful tools, but also easily accessible. In addition, [2] gauges the number of worldwide Excel users at more than 750 million, and Forrester Research measures 50 to 81% of businesses use spreadsheets [3]. Moreover, there is a

huge amount of data on the web that is only available via spreadsheets. For instance, the United States government published a compilation of thousands of spreadsheets about economic development, transportation, public health, and other important social topics; a spreadsheet was the only data format used.

Spreadsheets contain data that is viewed in a tabular form called semi-structured data. It uses tags, markers, standard structures, or hierarchal structure to encode the semantic units.

Thus, it is utilized by a huge number of clients as a standard generally useful data management tool. It is currently increasingly necessary for external applications and services to consume spreadsheet data. It is consisting of data regulated during a two-dimensional (2-D) cell. Normally, fragments during a table address the table name, and each line addresses the attributes or the values. The primary line or lines may contain segment attributes or variable names. Cells may contain numbers, text, dates, and other data types, or they will contain a formula (out of scope), that performs checks that reference the estimations of different cells. Past this fundamental model, Spreadsheets are utilized by an enormous number of clients as a typical generally useful data management tool. Since the concept of semi-structured data is given relative to the structured data, it is presently progressively essential for outdoor applications and administrations to consume spreadsheet data.

Unfortunately, the data stored in the spreadsheet is low-quality. In this thesis, we examine the issue of spreadsheets low-quality besides the importance need for changing over arbitrary tables in spreadsheets into structured formats required by the applications and services. The thesis

proposes a novel methodology in which extraction logic is embedded in a spreadsheet conversion process.

The issues of low quality for a spreadsheet whose object is data stockpiling are divided into two primary parts. The initial segment originates from the perspective on the spreadsheet itself, for example, multiple worksheets, hidden content, and various tables in the same sheet. In addition, the spreadsheet may contain tables and text, and diagrams on the same sheet. And so on. The subsequent part is explicit to the data stored in a spreadsheet in particularly the table content of the Excel spreadsheet. For example, blank cell or empty cell, no primary key, different formats: like distinctive date organizes in Excel spreadsheets, diverse text style or shading and duplicated data.

The increasing interest in ontology and semantics in recent years has led to the creation or use of ontology for different purposes and with different feature systems. We provide an approach that uses the database schema as an ontology schema supported to the source code to improve the quality of data stored in the spreadsheet tables such as duplicated recognition. This ontology technique can directly be executed with the Data Quality Management (DQM) process and extracted its knowledge representation as an output. With the proposed approach, high quality can be achieved when changing over spreadsheet data with low-quality to the relational model.

Metadata is a collection of key information about particular content, which can be used to facilitate the understanding, use, and management of data. The thesis focuses only on how to extract the table stores in a spreadsheet, not on graphs or text, or anything else. As a result, in this thesis, the metadata extractor acts to extract the semantic regions from the