FOSSEE Summer Fellowship Report

on

FLOSS - R

submitted by

Debatosh Chakraborty (National Institute of Technology, Agartala)

under the guidance of

Prof. Kannan M. Moudgalya
Chemical Engineering Department,
IIT Bombay

Prof. Radhendushka Srivastava
Mathematics Department,
IIT Bombay

and supervision of

Digvijay Singh
Project Research Associate,
R Team, FOSSEE,
IIT Bombay

August 16, 2022

Acknowledgment

I thank Prof. Radhendushka Srivastava, Department of Mathematics, IIT Bombay, for his guidance and suggestions throughout the project. I also express my sincere gratitude to Prof. Kannan M. Moudgalya, Department of Chemical Engineering, IIT Bombay, for creating the FOSSEE Summer Fellowship program and providing students from all over India an opportunity to participate in it. I equally respect the guidance provided by Mr. Digvijay Singh throughout the project.

Contents

Chapter 1

Introduction

In this report, I mention my contribution to open-source software (FLOSS) made during the Summer Internship, starting from 16th May 2022 to 16th August 2022. Contributions are made using a (Free-Libre/Open Source Software) known as ‘R’ as a part of the FOSSEE Project by IIT Bombay and MHRD, Government of India. The FOSSEE project is a part of the National Mission on Education through ICT. The thrust area is promoting and creating open-source software equivalent to proprietary software, funded by MoE, based at the Indian Institute of Technology Bombay (IITB). My contributions involve the creation of an R code to perform data manipulation for primary key generation and the creation of an R TBC.

Chapter 2

Contribution to the TBC project

As a part of the selection procedure for the FOSSEE Summer Fellowship, an applicant is required to select a standard textbook related to Probability, Statistics, Algebra, etc., with at least 80 solved examples to submit a TBC proposal for the R TBC project. My proposal got approved, and during the fellowship period, I contributed to the R TBC project by creating an R textbook companion for the below-mentioned textbook:

Table 1. Details of the textbook selected for R TBC contribution.

Textbook Name	**Author**	**Edition**
Mathematical Statistics and Data Analysis	John A. Rice	3rd

My submitted TBC shall be available for public use on the R TBC Completed Books webpage upon approval.

Chapter 3

Data manipulation for primary key generation using R

I was given the task to first find whether a method exists which could convert a dataset with multiple entries for a potential primary key into a matrix with only a single record for each distinct potential primary key value and if no such method exists then to help create it in R. The process involved conducting an exhaustive literature survey, assisting in the creation of an algorithm to solve the problem, and implementing it in the R programming language.

1. Introduction

Usually, the datasets involving customers' records are often very messy and sometimes devoid of a primary key. They involve multiple records corresponding to a particular unique id. For example, customer purchase history at different times may contain different purchase information, such as product name, quantity, price, etc., for the same customer. All information about a particular customer is maintained as separate records in the dataset.

The idea was to restructure the dataset in such a way that all the information of a particular customer is present in a single row. For a sample customer purchase history dataset, as shown below:

Table 2. Sample customer purchase history dataset.

Customer_id	**Product_id**	**Price**	**Discount**	**Quantity**
1	20032	300	10	1
1	20032	300	20	3
1	20032	300	50	5
1	20035	100	20	1
1	20035	100	20	2

The transformed dataset should look as follows:

Table 3. Transformed customer purchase history dataset.

Customer_id	Product_id.1	Product_id.2	Price.1	Price.2	Discount.1	Discount.1
1	20032	20035	300	100	10	20

Discount.3	Quantity.1	Quantity.2	Quantity.3	Quantity.4
50	1	3	5	2

2. Literature Survey

Under the guidance of Prof. Radhendushka Srivastava, I performed an exhaustive literature survey to find any existing solution to the data transformation problem. I searched for customer database handling solutions, general data transformation tools (both proprietary and open source), R packages, and various publications on data transformation. Following is a list containing all search results:

Table 4. Search results for data transformation tools and publications.

S. No.	Title	Year	Author	Publisher
1	shinyplyr	2020	David Barke	
2	Analysis and R shiny application on eCommerce data	2019	Qifan Wang	NYC Data Science
3	Open Refine	2012	Google	
4	Trifecta		Trifecta	
5	Wrangler: Interactive Visual Specification of Data Transformation Scripts	2011	Sean Kandel, Andreas Paepcke, Joseph Hellerstein, Jeffrey Heer	Standford

6	The R Language as a Tool for Biometeorological Research	2020	Ioannis Charalampopoul	Atmosphere
7	From 5Vs to 6Cs: Operationalizing Epidemic Data Management with COVID-19 Surveillance	2020	Akhil Sai Peddireddy; Dawen Xie; Pramod Patil; Mandy L. Wilson; Dustin Machi; Srinivasan Venkatramanan	IEEE
8	Automating Data Preparation: Can We? Should We? Must We?	2019	Norman Paton	21st International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data
9	Fundamentals of Wrangling Healthcare Data with R	2022	Wickham and Grolemund 2017	
10	Towards Automatic Data Format Transformations: Data Wrangling at Scale	2017	Alex Bogatu(B), Norman W. Paton, and Alvaro A.A. Fernandes	British International Conference on Databases

No relevant tool or code was found during the search. Tools to perform data transformations like filtering, merging, removing, and transposing were found, but no tool was found which could generate a primary key from a column of the input dataset.

After analyzing the search results, Prof. Radhendushka Srivastava advised me to look for research literature in the domain of data transformation. He suggested searching for popular datasets where the data format is similar to the one illustrated in Table 2 and listing all research articles related to their transformation and analysis. Various recommender system datasets match the data format of Table 2; hence I started searching research literature associated with eight different popular recommender system datasets, namely, MovieLens, Million Songs, Netflix, Steam Video Games, Amazon, Books Crossing, LastFM, and Free Music Archive.

I made use of the following keywords on Google Scholar to search for relevant articles:

1. dataset_name exploratory data analysis
2. dataset_name preprocessing
3. dataset_name data wrangling
4. dataset_name transformation
5. dataset_name reformatting
6. dataset_name primary key creation
7. dataset_name cleaning
8. dataset_name rdbms creation
9. dataset_name relational form
10. dataset_name conversion to data matrix
11. dataset_name creation of user matrix
12. dataset_name matrix factorization
13. dataset_name dimension reduction
14. recommendation system analysis
15. recommendation system data transformation and preprocessing
16. recommendation system data transformation and preprocessing

In the above list, the term **dataset_name** is replaced with the name of the recommender system dataset before searching. Out of all the keywords mentioned above, the first seven yielded the most relevant results.

I went through the abstract and conclusion of the search results to check their relevance. For relevant results, I went through the content of their data transformation and analysis section to better understand the tools and techniques mentioned for data transformation. After presenting the final list of search results to Prof. Radhendushka Srivastava and Mr. Digvijay Singh, I removed some irrelevant items as per their suggestions. The final list is shown below:

Table 5. Search results for research literature on recommender system datasets.

Topic	S.No.	Title	Link
Data Transformation Tools	1	Converting heterogeneous statistical tables on the web to searchable databases	https://doi.org/10.1007/s10032-016-0259-1
	2	TabbyXL: Rule-Based Spreadsheet Data Extraction and Transformation	10.1007/978-3-030-30275-7_6

	3	Data Preparation: A Survey of Commercial Tools	https://doi.org/10.1145/3444831.3444835
	4	Foofah: Transforming Data By Example	https://doi.org/10.1145/3035918.3064034
	5	RDF123: From Spreadsheets to RDF	10.1007/978-3-540-88564-1_29
MovieLens Dataset	6	iSynchronizer: A Tool for Extracting, Integration and Analysis of MovieLens and IMDb Datasets	https://doi.org/10.1145/3213586.3226219
	7	Movie Dataset Analysis Using Hadoop-Hive	https://doi.org/10.1109/CSITSS.2017.8447828
	8	Extraction and Integration of MovieLens and IMDb Data	Microsoft Word - TR-ExtractionIntegration-21.doc (uvsq.fr)
Million Songs Dataset	9	A Preliminary Study on a Recommender System for the Million Songs Dataset Challenge	(PDF) A Preliminary Study on a Recommender System for the Million Songs Dataset Challenge (researchgate.net)

	10	F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4: 19:1–19:19.	https://doi.org/10.1145/2827872
	11	Million Song Dataset	The Million Song Dataset | Academic Commons (columbia.edu)
	12	The MovieLens Datasets: History and Context	The MovieLens Datasets: History and Context: ACM Transactions on Interactive Intelligent Systems: Vol 5, No 4
	13	Music Recommender System CS365: Artificial Intelligence	report.pdf (iitk.ac.in)
	14	Variational Autoencoders for Collaborative Filtering	Variational Autoencoders for Collaborative Filtering | Proceedings of the 2018 World Wide Web Conference (acm.org)
	15	Embarrassingly Shallow Autoencoders for Sparse Data	Embarrassingly Shallow Autoencoders for Sparse Data (arxiv.org)

	16	Ontology-based Recommender System for the Million Song Dataset Challenge	Ontology-based Recommender System for the Million Song Dataset Challenge | IEEE Conference Publication | IEEE Xplore
Book-Crossing Dataset	17	Hybrid attribute and personality based recommender system for book recommendation	Hybrid attribute and personality based recommender system for book recommendation | IEEE Conference Publication | IEEE Xplore
	18	Study of Distributed Framework Hadoop and Overview of Machine Learning using Apache Mahout	https://doi.org/10.1109/CCWC.2019.8666529
	19	Introducing Hybrid Technique for Optimization of Book Recommender System	https://doi.org/10.1016/j.procs.2015.03.075
Amazon Review Dataset	20	Sentiment analysis on large scale Amazon product reviews	Sentiment analysis on large scale Amazon product reviews | IEEE Conference Publication | IEEE Xplore

	21	Amazon review analysis	Amazon-Reviews-Sentiment-Analysis-A-Reinforcement-Learning-Approach.pdf (researchgate.net)
	22	EDA on Amazon Data	International Journal of Innovative Technology and Exploring Engineering (IJITEE) (researchgate.net)
Yahoo Music User Ratings Dataset	23	Leakage in Data Mining	Leakage in data mining: Formulation, detection, and avoidance: ACM Transactions on Knowledge Discovery from Data: Vol 6, No 4
	24	Big Data Frameworks for Sites and Products Recommendation	Big Data Frameworks for Sites and Products Recommendation | Journal of Information (conscientiabeam.com)
	25	Yahoo! music recommendations: modeling music ratings with temporal dynamics and item taxonomy	Yahoo! music recommendations | Proceedings of the fifth ACM conference on Recommender systems

LastFM Dataset	26	Inclusion of Semantic and Time-Variant Information Using Matrix Factorization Approach for Implicit Rating of Last.Fm Dataset	Inclusion of Semantic and Time-Variant Information Using Matrix Factorization Approach for Implicit Rating of Last.Fm Dataset | SpringerLink
	27	MUSIC MOOD DATASET CREATION BASED ON LAST.FM TAGS	Microsoft Word - 03. AIAP 01 (csitcp.com)
	28	All You Need is Ratings: A Clustering Approach to Synthetic Rating Datasets Generation	[1909.00687] All You Need is Ratings: A Clustering Approach to Synthetic Rating Datasets Generation (arxiv.org)
Steam Video Games Dataset	29	Recommender Systems for Online Video Game Platforms: the Case of STEAM	Recommender Systems for Online Video Game Platforms: the Case of STEAM | Companion Proceedings of The 2019 World Wide Web Conference (acm.org)
	30	Game Achievement Analysis: Process Mining Approach	Game Achievement Analysis: Process Mining Approach | SpringerLink
	31	Hybrid system for video game recommendation based on	Hybrid system for video game

		implicit ratings and social networks	recommendation based on implicit ratings and social networks | SpringerLink
Free Music Archive	32	Automatic Music Production Using Generative Adversarial Networks	Automatic Music Production Using Generative Adversarial Networks | OpenReview
	33	The becoming of an archive: perspectives on a music archive and the limits of institutionality	The becoming of an archive: perspectives on a music archive and the limits of institutionality: Social Dynamics: Vol 46, No 2 (tandfonline.com)
	34	A Novel Approach for Music Recommendation System Using Matrix Factorization Technique	A Novel Approach for Music Recommendation System Using Matrix Factorization Technique | SpringerLink

No tool or algorithm was found during the search which could transform a dataset in the way which we require. Hence we went ahead with the construction of an algorithm to solve the problem.

3. Creation of algorithm for data transformation

For the purpose of transforming a dataset into a matrix with only a single record for each distinct potential primary key/unique id value, I proposed a column transformation approach. In this approach, all the columns of the original dataset are separately processed. For a particular column, all the unique values associated with a potential primary key column value are found and broken into multiple columns under the same name but with sequential numbering based on the occurrence of a value under that column header. This process is repeated for all potential primary key column values. Once all the columns are processed, they are merged into a single entity.

The algorithm for this approach is described below:

Step 1: Create pairs of the potential primary key column with the rest of the dataset columns.

Step 2: For each pair, follow the below-mentioned steps:

- **a.** Obtain distinct entries.
- **b.** Create a new column with the name **count** containing the sequential numbering of the data entries for each potential primary key column value.
- **c.** Execute the **reshape()** function of R by passing the potential primary key column name to its **idvar** argument, **count** column to its **timevar** argument, and the paired data column to its **v.names** argument.

Step 3: Bind all the columns.

R code to implement the algorithm:

1. **shape():** This function implements steps 2(b) and 2(c) of the algorithm.

```
shape = function(data, pr, col){ # "data" is the input dataset, "pr" is the potential primary key, "col" is data column

# Filtering data to be restructured
  data = data[,c(pr, col)]

# Step 2(b) of the algorithm
  temp = data %>%
    unique() %>%
    cbind(index = 1:nrow(.),.) %>%
    cbind(., count = ave(.$index, .[pr], FUN = seq_along))

# Step 2(c) of the algorithm
  matrix = reshape(temp[, names(temp) != "index"],  idvar = pr, timevar = "count", v.names = col,
direction = "wide")
  matrix[,-1]
}
```

2. **transform():** This function implements the complete algorithm by incorporating the **shape()** function in itself.

```
transform = function(data, col_nm = colnames(data), pr_key){# "data" is the input dataset, "col_nm" contains all column headers of the input dataset, and "pr_key" is the potential primary key

# If "col_nm" contains "pr_key" then remove it
col = col_nm[col_nm != pr_key]
```

```r
# Steps 1 and 2 of the algorithm
data_col = lapply(col, shape, data = data, pr = pr_key)
final_data = unique(data[,c(pr_key)])


# Step 3 of the algorithm
for (i in 1:length(data_col)) {
  final_data = cbind(final_data, data_col[[i]])
  colnames(final_data)[1] = pr_key
  final_data
}
```

Chapter 4

Conclusion

The projects completed during the FOSSEE Summer fellowship contributed towards the increment in usability and awareness of open-source software, i.e., R. Completed R TBCs are made available to the general public to be used as a companion to the associated standard textbooks in Mathematics and Sciences. The data manipulation project demonstrates R's capabilities in customizing data for a specific use case.

Overall, it was a great learning experience. I gained new skills and knowledge. I also learned the different facets of working within an organization. In a nutshell, the fellowship taught me work ethics, commitment, and the importance of contributing back to society besides technical skills.