ISSN 2278-3091
Volume 9, No.5, September - October 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse58952020.pdf
https://doi.org/10.30534/ijatcse/2020/58952020

Data Integration and Database Formation of Historical Geographical Datasets

Sameer Kaul[1], Majid Zaman[2], Muheet Ahmed[3]
[1] Research Scholar, Department of Computer Science, University of Kashmir, kaulsameer78@gmail.com
[2] Scientist, Directorate of Information Technology & Support System, University of Kashmir, India
zamanmajid@gmail.com
[3]Scientist, Department of computer sciences, University of Kashmir, India

ABSTRACT

Knowledge discovery from the geographical dataset is critical for understanding weather behavior and global warming impact. However Geographical data is complex and growing exponentially. A major effort goes into the accusation of geographical (raw) data, which contains possible errors, is un-validated, unformatted, un-coded with missing and wrong values. Resultantly Geographical database development costs up to 70% [1] (or more) of time and effort. In this paper, we present novel algorithms to convert raw geographical data into error-free, clean, validated and formatted Databases. Further resultant geographical database is visualized for further understanding of dataset, which shall works as impetus for data engineers and scientists [3][4][5][10-15].

Key words: geographical data, data extraction, data integration, data pre-processing, database, data visualization.

1. INTRODUCTION

Geographical data is growing exponentially, massive volumes of geographical data are generated having multiple columns and most of the times stored in comma-separated values (CSV) files. However, not all relevant generated data is stored in single file but may result in multiple files. While CSV files are simple means of storing and retrieving data, the challenge lies in analyzing data stored in CSV files especially when data is stored in multiple files. Further, geographical data is generated incrementally and CSV files cannot be considered as means of permanent storage. Database provides solution to both the problems, one of storing data permanently and incrementally besides having to access data in a rich, uniform manner, e.g. using Structured Query Language (SQL) would offer expressiveness and user-friendliness [2][6][7][8][16-20].

2. DATASET

In this study weather prediction for Kashmir region has been performed using data mining techniques. The datasets used here has been collected from NDC Pune (India Meteorological department). It is an agency of Ministry of earth sciences of the government of India. It is the principal agency responsible for meteorological observations, weather forecasting and seismology. IMD is one of the six regional specialized meteorological centers of the world meteorological organization. The first dataset used consists of relative humidity (Fig 2) measured (in %) from year 2012 to 2017 measured every day at time 3 PM and 12 AM. The dataset consists of 12190 instances and five dimensions. The second dataset used consists of various weather parameters measured every day from year 2012 to 2017.Dataset consists of 6117 instances and five dimensions. The parameters included in the dataset are Date, Maximum temperature (°C), Minimum temperature (°C) and Rainfall (in mm) as shown in Fig 1. The parameters recorded in the dataset depict that some months like October, November and December shows less rainfall than rest of the months in all the regions. The weather parameters in both datasets are taken for the 3 regions of Kashmir division i.e. Gulmarg (North Kashmir), Srinagar (Central Kashmir) and Qazigund (South Kashmir). The geographical description of these areas specifies that Gulmarg is located at 34.05°N 74.38°E [1] and has an average elevation of 2,650 m (8,690 ft), Srinagar (Central) is located at 34.5°N 74.47°E [1] and has an average elevation of 1,585 m (5,200 ft), and Qazigund (South) is located at 33.59°N 75.16°E [1]. It has an average elevation of 1,670 m (5,480 ft) [1][9].

station_id	year	mnth	dt	tmax	tmin	rfall
42026	2012	1	1	5.5	-8	0
42026	2012	1	2	5.4	-7.6	0
42026	2012	1	3	4.2	-8	0
42026	2012	1	4	4	-7.2	0
42026	2012	1	5	-1	-9.1	1.1
42026	2012	1	6	-2	-8	17.9
42026	2012	1	7	-1	-10.5	6.8
42026	2012	1	8	1	-16.5	12.6
42026	2012	1	9	-2.8	-14.5	0
42026	2012	1	10	-2.5	-16.2	0
42026	2012	1	11	-7.8	-14.8	0
42026	2012	1	12	-8.2	-16.4	0
42026	2012	1	13	-7.5	-16.5	0
42026	2012	1	14	-7.5	-15.2	0
42026	2012	1	15	-1.5	-9.6	16
42026	2012	1	16	-3	-6.7	21

Figure 1: Instances of Maximum Temperature, Minimum

Temperature and Rainfall

station_id	year	mnth	hr	dt	rhumid
42026	2012	1	3	1	100
42026	2012	1	3	2	100
42026	2012	1	3	3	96
42026	2012	1	3	4	100
42026	2012	1	3	5	100
42026	2012	1	3	6	100
42026	2012	1	3	7	100
42026	2012	1	3	8	100
42026	2012	1	3	9	100
42026	2012	1	3	10	86
42026	2012	1	3	11	87
42026	2012	1	3	12	100
42026	2012	1	3	13	100
42026	2012	1	3	14	100
42026	2012	1	3	15	100
42026	2012	1	3	16	100
42026	2012	1	3	17	100

Figure 2: Instances of Relative Humidity

station_id	year	mnth	dt	tmax	tmin	rfall	humid3	humid12
42026	2012	1	1	5.5	-8	0	100	100
42026	2012	1	2	5.4	-7.6	0	100	100
42026	2012	1	3	4.2	-8	0	96	90
42026	2012	1	4	4	-7.2	0	100	100
42026	2012	1	5	-1	-9.1	1.1	100	100
42027	2012	1	1	10.5	-4.9	0	92	53
42027	2012	1	2	10.5	-3.6	0	92	57
42027	2012	1	3	10.6	-2.5	0	81	61
42027	2012	1	4	11	-3.1	0	89	73
42027	2012	1	5	4.8	0.7	0.2	93	85
42044	2012	1	1	10.5	-5	0	84	55
42044	2012	1	2	8.5	-1.6	0	90	60
42044	2012	1	3	10.5	-3	0	96	56
42044	2012	1	4	11.2	-3.4	0	85	67
42044	2012	1	5	7	0	4.1	90	93

Figure 3 : Instances of Integrated Data

3. DATA STRUCTURE

To carry out the analysis of the meteorological data set, three N*M dimensional data structures are used, details of such data structures are as follows:

3.1. Humidity (hu_data):

The Humidity data table (hu_data) contains the following fields: station_id, year, mnth, hr, dt, rhumid. The station_id field specifies a particular station of Kashmir from which readings are recorded. year, mnth, dt fields specifies a particular date i.e. day, month and year on which readings of the station are recorded, hr field in the table implies the hour in the day at which readings are recorded (twice a day at 12AM and 3PM) and the last field rhumid specifies the percentage of humidity recorded from a particular station on a particular day at 12 AM or 3 PM. The table structure is given below Fig 4:

Field	Type	Null	Key	Default	Extra
station_id	int(11)	YES		NULL	
year	int(11)	YES		NULL	
mnth	int(11)	YES		NULL	
hr	int(11)	YES		NULL	
dt	int(11)	YES		NULL	
rhumid	int(11)	YES		NULL	

Figure 4: Relative Humidity Data Structure

3.2. Rainfall data table (rn_data):

The rainfall data table (rn_data) contains the following fields: station_id, year, mnth, dt, tmax, tmin, rfall. The station_id field specifies a particular station from which readings are recorded. Year, mnth, dt fields specifies a particular date i.e. Day, month and year on which readings of the station are recorded, tmax and tmin fields specifies the maximum and minimum temperatures respectively recorded from a station in degree Celsius (°c) for a particular day. Rfall field specifies the rainfall recorded by the station on a particular day. The table structure is given below: Fig 5

Field	Type	Null	Key	Default	Extra
station_id	int(11)	YES		NULL	
year	int(11)	YES		NULL	
mnth	int(11)	YES		NULL	
dt	int(11)	YES		NULL	
tmax	float	YES		NULL	
tmin	float	YES		NULL	
rfall	float	YES		NULL	

Figure 5: Rainfall Data Structure

3.3. Combined data table (C_DATA):

The combined data table (c_data) contains the following fields: station_id, year, mnth, dt, tmax, tmin, rfall, humid3, humid12. It is formed by adding the humid3 and humid12 fields to the rn_data table, where humid3 and humid12 fields specify the humidity readings recorded by a station on a particular at 3 pm and 12 am respectively. The table structure is given below: Fig 6

Field	Type	Null	Key	Default	Extra
station_id	int(11)	YES		NULL	
year	int(11)	YES		NULL	
mnth	int(11)	YES		NULL	
dt	int(11)	YES		NULL	
tmax	float	YES		NULL	
tmin	float	YES		NULL	
rfall	float	YES		NULL	
humid3	int(11)	YES		NULL	
humid12	int(11)	YES		NULL	

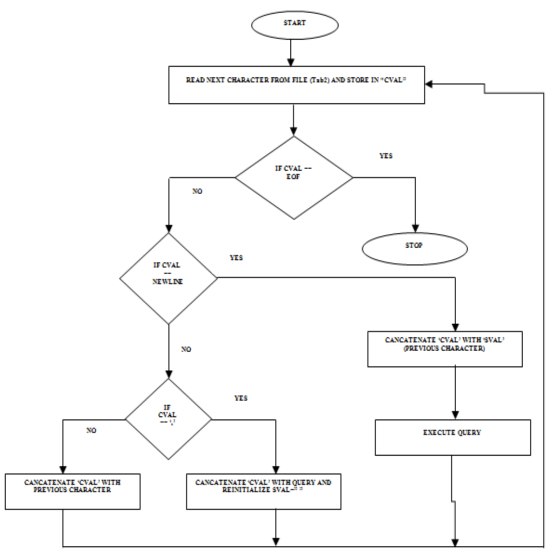
Figure 6: Combined Data Structure.

4. IMPLEMENTATION

4.1. Algorithm 1

The algorithms reads a character from comma separated file (csv) until end of file is reached, if the current character is newline the whole row is inserted in the “rn_data” and “hu_data” tables otherwise if the character is “,” (comma) or any value, it is concatenated with previous value. Below two flowcharts specifies how the data is inserted in “rn_data” and “hu_data” tables from files (tab2 and tab3) respectively:

Flowchart 1:

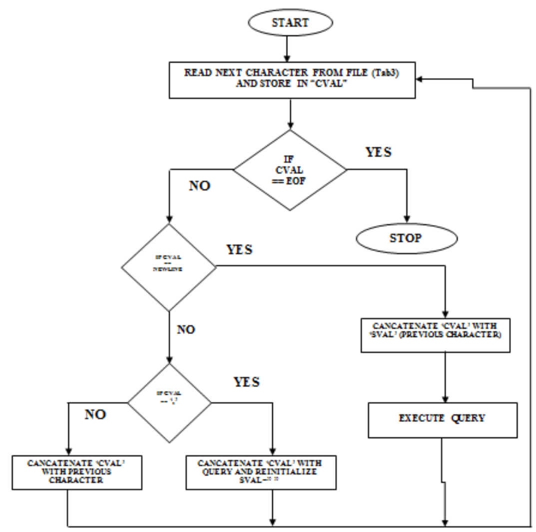


Working:

Declare
  Set sval = “ ”
  Set i = 0
  Set avl = Read Character from file (Tab2)
  Set cval = avl
Set qry = “ ”

Begin
  for i = 0 to avl by 1 do
    Set avl = Read Next character from file
    If cval = “Newline” then
      Set sval = sval + cval // Concatenate cval with sval
      Execute Query
    Endif
    If cval = “comma” then
      Set qry = qry + sval // concatenate cval with query
      Set sval = “ ” // Reinitialize sval
    Elseif
      Set sval = sval + cval // concatenate cval with previous character
    Endif
  Endfor
End

Flowchart 2:



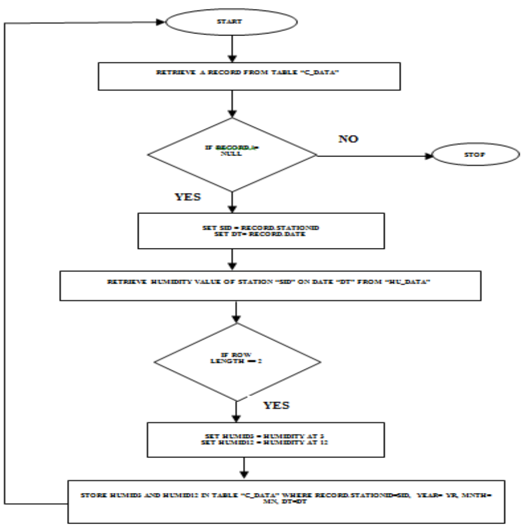
Working:

Declare
  Set sval = “ ”
    Set i = 0
    Set avl = Read Character from file (Tab3)
    Set cval = avl
      Set qry = “ ”
Begin
  for i = 0 to avl by 1 do
    Set avl = Read Next character from file
    If cval = “Newline” then
      Set sval = sval + cval // Concatenate cval with sval
      Execute Query
    Endif
    If cval = “comma” then
      Set qry = qry + sval // concatenate cval with query
      Set sval = “ ” // Reinitialize sval
    Elseif
      Set sval = sval + cval // concatenate cval with previous character
    Endif
  Endfor
End

4.2. Algorithm 2

We are given a table named "c_data" consists of data recorded every day from each station and other table named "hu_data" consists of humidity readings of each station recorded every day at 12 AM and 3 PM. The algorithm retrieves the humidity readings of each station whose data resides in table "c_data". These humidity readings consist of two values for each station on respective dates, one reading taken at 12 AM and other at 3 PM. After retrieving the humidity readings, the algorithm stores these values in the table "c_data" along with the respective stations on the respective dates.

Flowchart:



Working:

Declare
Set RS = " "
Set SID= " "
Set DT= " "
Set Yr **=** " "
Set DT = " "
Set MNTH = " "
Begin
Set RS = Read data from table (**C_DATA**) // Retrieve data from C_DATA
While RS != NULL **do**
Set SID = RS.STATION_ID
Set DT = RS.DATE
Retrieve Humidity value from **Hu_Data** table **Where**
STATION_ID = "SID" **&&**
YEAR = "YR" **&&**
MONTH = "MNTH" **&&**
DATE = "DT"

If ROW_LENGTH = 2 **Then**
Set HUMID3 = HUMIDITY at 3
Set HUMID12 = HUMIDITY at 12
Endif
Set HUMID3 in C_DATA **&&**
Set HUMID12 in C_DATA **where**
RS.STATION_ID = SID**,**
YEAR = "YR" **&&**
MNTH = "MNTH" **&&**
DT = "DT"

Execute Query

Endwhile
End

5. 3D Visualization

Geographical data is integrated and subsequently database is populated. 3D visualization is performed on geographical database primarily on following attributes:

1. Minimum Temperature
2. Maximum Temperature
3. Humidity at 12 A.M
4. Humidity at 3 P.M
5. Rain Fall

and the resultant output is shown below: (Fig 7 – 10)

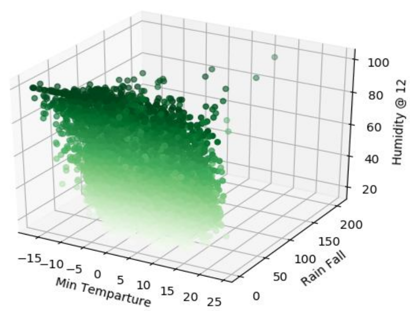


Figure 7: Rain Fall, Minimum Temperature & Humidity at 12 AM

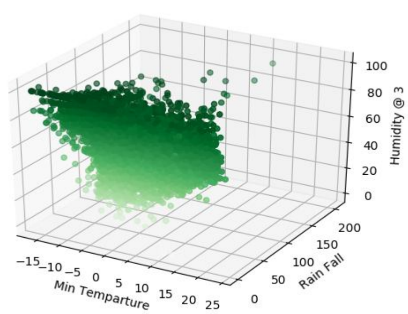


Figure 8: Rain Fall, Minimum Temperature & Humidity at 3 PM

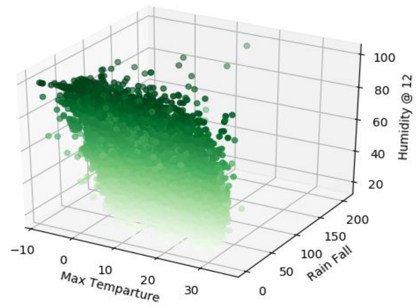


Figure 9: Rain Fall, Maximum Temperature & Humidity at 12 AM

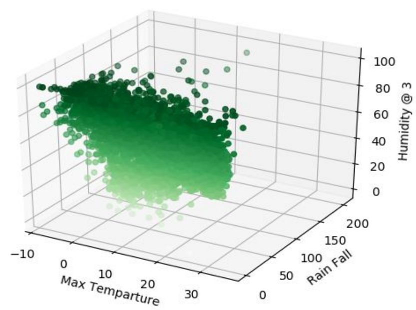


Figure 10: Rain Fall, Maximum Temperature& Humidity at 3 PM

6. CONCLUSION

This paper features how meaningful knowledge can be extracted from raw geographical dataset. The proposed algorithm extracts, transforms & integrates data from multiple CSV files and finally populates formulated database, resultant database is clean, integrated and ready for analysis. In this paper 3D visualization is done to further understand geographical data. Finally integrated geographical database is ready for analysis and machine learning.

REFERENCES

1. Ömer Mutluoglu, Ayhan Ceylan. **“Accuracy and cost comparison of spatial data- acquisition methods for the development of geographical information systems”.** *Journal of Geography and Regional Planning* Vol. 2(9), pp. 235-242, September, 2009.
2. Adaszewski S (2014) Mynodbcsv: **Lightweight Zero-Config Database Solution for Handling Very Large CSV Files. PLOS ONE** 9(7):e103319. https://doi.org/10.1371/journal.pone.0103319
3. Ristoski, Petar, and Heiko Paulheim. **"Semantic Web in data mining and knowledge discovery: A comprehensive survey."** *Journal of Web Semantics* 36 (2016): 1-22.
4. Miller, H. (Ed.), Han, J. (Ed.). (2009**). Geographic Data Mining and Knowledge Discovery. Boca Raton**: CRC Press, https://doi.org/10.1201/9781420073980
5. Fan, W., Lu, H. Madnick, S.E., Cheung, D. (2001), **Discovering and reconciling value conflicts for numerical data integration. Information Systems**. 26:635-656.
6. Fayyad, U.M., Piatetsky-Shapiro, G. Smyth, P. (1996), from data mining to knowledge discovery: An Overview. In Fayyad, U.M., Piatetsky-Shapiro, G. Smyth, P. Ulthurusamy,R. (eds) Advances in Knowledge Discovery and Data Mining. Cambridge, MA:MIT Press, 1-34.
7. Zhang, Chao, and Jiawei Han. **"Geographic Data Mining." International Encyclopedia of Geography: People, the Earth, Environment and Technology: People, the Earth, Environment and Technology** (2016): 1-5.
8. Papoušek, Jan, Radek Pelánek, and Vít Stanislav. "**Adaptive geography practice data set."** *Journal of Learning Analytics* 3.2 (2016): 317-321.
9. Majid Zaman, Sameer Kaul and Muheet Ahmed. **―Analytical Comparison between the Information Gain and Gini Index using Historical Geographical Data‖**. *(IJACSA) International Journal of Advanced Computer Science and Applications*, (pp. 429-440), Vol. 11, No. 5, 2020.
10. Zaman, Majid, S. M. K. Quadri, and Muheet Ahmed Butt. **"Information Translation: A Practitioners Approach."** In Proceedings of the World Congress on Engineering and Computer Science, vol. 1. 2012.
11. Mohd, Razeef, Muheet Ahmed Butt, and Majid Zaman Baba. **"GWLM–NARX."** Data Technologies and Applications (2020).
12. Ashraf, Mudasir, Majid Zaman, and Muheet Ahmed. **"Performance analysis and different subject combinations: An empirical and analytical discourse of educational data mining."** In 2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence), pp. 287-292. IEEE, 2018.
13. Ashraf, M., Zaman, M. and Ahmed, M., 2020. **An Intelligent Prediction System for Educational Data Mining Based on Ensemble and Filtering approaches. Procedia Computer Science**, 167, pp.1471-1483.
14. Ashraf, Mudasir, Majid Zaman, and Muheet Ahmed. **"To ameliorate classification accuracy using ensemble vote approach and base classifiers."** In Emerging Technologies in Data Mining and Information Security, pp. 321-334. Springer, Singapore, 2019.
15. Mir, Nasir Majeed, Sarfraz Khan, Muheet Ahmed Butt, and Majid Zaman. **"An experimental evaluation of bayesian classifiers applied to intrusion detection."** Indian Journal of Science and Technology 9, no. 12 (2016): 1-7.
16. Ashraf, Mudasir, Syed Mudasir Ahmad, Nazir Ahmad Ganai, Riaz Ahmad Shah, Majid Zaman, Sameer Ahmad Khan, and Aftab Aalam Shah. "Prediction of Cardiovascular Disease through Cutting-Edge Deep Learning Technologies: An Empirical Study Based on TENSORFLOW, PYTORCH and KERAS." In

International Conference on Innovative Computing and Communications, pp. 239-255. Springer, Singapore.

17. Ashraf, Mudasir, Majid Zaman, and Muheet Ahmed. "Using Ensemble StackingC Method and Base Classifiers to Ameliorate Prediction Accuracy of Pedagogical Data." Procedia computer science 132 (2018): 1021-1040.
18. Mohd, Razeef, Muheet Ahmed Butt, and Majid Zaman Baba. "SALM-NARX: Self Adaptive LM-based NARX model for the prediction of rainfall." In 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), 2018 2nd International Conference on, pp. 580-585. IEEE, 2018.
19. Putra, R. R., M. E. Johan, and E. R. Kaburuan. "A Naïve Bayes Sentiment Analysis for Fintech Mobile Application User Review in Indonesia." International Journal of Advanced Trends in Computer Science and Engineering 1860 (1856).
20. Syamala, M., and N. J. Nalini. "A Deep Analysis on Aspect based Sentiment Text Classification Approaches." International Journal of Advanced Trends in Computer Science and Engineering 8, no. 5 (2019): 1795-1801.