# Big Data Domain Research in Government Domain

**Abstract**

The domain, which I chose to analyze data for my project, is **Urban Planning**, which is a subsidiary domain of **Government** domain. This domain mainly deals with land usage statistics in towns and cities of United Kingdom. Mainly this land use is characterized into nine categories like domestic buildings, non-domestic buildings, roads, paths, rail, gardens (domestic), green space, water, other land uses (largely hard standing), and unclassified. This data facilitates analysis of housing density in a particular geographic area and thus providing data for where to build new housing in an area. The main problem here is to give a clear picture of how much land is being used for residential purposes and how much is being left for green spaces in a particular geographic area. One more problem, which I identified in this domain, is, how weather in a particular geographic region is affected by the land utilization in that area. Collecting weather data along with land usage statistical data will provide a good framework for data analysis and to provide better solutions for the problems.

**Data Sources, Preparing the data and Analysis**

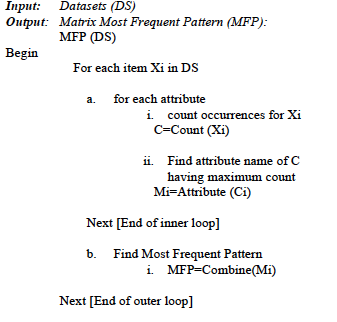
First, I will introduce to the procedure of how data is collected and analyzed from data sources for analysis and later on will move on how I collected my data for the project. First of all, we need to know -

**What is Big Data Analytics?**

Big data analytics is the process of examining large data sets containing a variety of data types to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits. So, the first step in data analysis is finding of human interpretable patterns in the data collected. For finding these patterns we can make use of a data-mining algorithm such as Most Frequent Pattern (MFP) algorithm.

**MFP Algorithm**

Association rule mining is one of the most important and well defined technique for extract correlations, frequent patterns, associations or causal structures among sets of items in the transaction databases or other repositories. The MFP algorithm deals with generation of frequent patterns and strong association between them. For this purpose a property matrix containing counted values of corresponding properties of each product has been used as shown in the figure below. Let we have set X of N items in a Dataset having set Y of attributes. This algorithm counts maximum of each attribute values for each item in the dataset. The MFP algorithm is as follows –



**Data for my project**

The main data source for my project is from the UK government data on land use statistics. After collecting this data, I observed that this data comprises of different regions in land use like Lower Layer Super Output Areas (LSOAs), Middle Layer Super Output Areas (MSOAs), Local Authorities (LAs), and Government Office Regions (GORs) in England in a particular year. All the land data collected is in square meters of area. I collected this data and identified human interpretable patterns in my data using MFP algorithm as follows –

**Step 1**

The first step in my data analysis is pre-processing of data collected. Data in its original format never confirm to the required shape for data mining. It needs to be transformed, integrated, and aggregated so that the mining process can effectively perform on it. I pre-processed my data and picked the key attributes in that data for illustration purposes. I picked these attributes, because these play a vital role in identifying human interpretable patterns. In this case the collected data in is cleaned by using SQL Server Data Transformation Services, and then removed noise from the transformed data. The transformed data is classified into three categories based on area of buildings, industrial buildings and green spaces as follows -

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **GOR\_Name** | **MSOA\_Name** | **Total Area (In m2)** | **Area of Buildings (In m2)** | **Area of Industrial Buildings (In m2)** | **Area of Green Spaces (In m2)** |
| North East | Chesterfield | 8075.52 | 2300.53 | 5000.01 | 775 |
| North East | Derbyshire | 13274.87 | 250.07 | 13000 | 24.80 |
| North West | Erewash | 3859.31 | 170.53 | 3400.22 | 389 |
| North West | High Peak | 1775.43 | 194.79 | 1400 | 180 |
| East Midlands | Blaby | 5820.06 | 217.93 | 5000.07 | 600.87 |
| East Midlands | Charnwood | 3573.27 | 245.53 | 3000.67 | 200.98 |

Above is the table with data containing high occupancy of industrial buildings and high green spaces.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **GOR\_Name** | **MSOA\_Name** | **Total Area (In m2)** | **Area of Buildings (In m2)** | **Area of Industrial Buildings (In m2)** | **Area of Green Spaces (In m2)** |
| North East | Hampshire | 5132.76 | 4900.45 | 130.31 | 100 |
| North East | Devon | 1131.43 | 1000 | 50.65 | 65.30 |
| North West | Treddyfrin | 3686.41 | 3200 | 250.43 | 200.32 |
| North West | Chester | 1431.08 | 900.65 | 289.67 | 76.23 |
| East Midlands | Gateshead | 1213.91 | 945.82 | 342.72 | 63.41 |
| East Midlands | New Castle | 1041.91 | 783.29 | 323.48 | 24.10 |

Above is the table with data containing high occupancy of domestic buildings and low green spaces.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **GOR\_Name** | **MSOA\_Name** | **Total Area (In m2)** | **Area of Buildings (In m2)** | **Area of Industrial Buildings (In m2)** | **Area of Green Spaces (In m2)** |
| Yorkshire | Uttlesford | 3843.96 | 2546.36 | 1325.78 | 12.76 |
| Yorkshire | Broxbourne | 418.69 | 327.45 | 125.67 | 10.65 |
| Humber | Dacorum | 1425.65 | 967.39 | 765.43 | 9.1 |
| Humber | Hertsmere | 652.9 | 432.76 | 289.37 | 3.4 |
| West Midlands | St Albans | 816.77 | 730.34 | 129.60 | 2.90 |
| West Midlands | Hertfordshire | 614.03 | 467.65 | 132.65 | 13.67 |

Above is the table with data containing high occupancy of domestic and industrial buildings with very low green spaces.

**Step 2**

The next after identifying patterns in our data is to build a MFP matrix with human interpretable patterns.

|  |
| --- |
| Most frequent patterns mined out of given data |
| 1. North East – Chesterfield – 5000.01 (Industrial Building) – 775 (Green Space) |
| 2. East Midlands – Blaby - 5000.01 (Industrial Building) – 600.87 (Green Space) |
| 3. North East – Devon – 50.65 (Industrial Building) - 65.30 (Green Space) |
| 4. West Midlands – Hertfordshire – 132.65 (Industrial Building) – 13.27 (Green Space) |

From above MFP matrix, I came to a conclusion that in a particular Government Office Region (GOR), in a Middle Layer Super Output Areas (MSOA), most of the green space is used up by domestic buildings. I observed that increase in domestic land utilization lead to decrease in green space areas.

The main problem in this domain that needs to be solved is how can we conserve green spaces and increase rainfall in a particular geographic region. So, for this I clustered the data collected using **K Means** algorithm.

**K Means Algorithm**

This is one of the widely known and implemented data-clustering algorithms. This is mainly used for classification of data. In this algorithm, the main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The steps of the algorithm are –

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.

The collected data is divided into clusters based on the K Means algorithm.

**Real Time Problems in this domain**

I will quote two real time problems in the government domain. I will explain the scenarios of problems and what is the proposed solution for the problems.

**Metro Rail Project for a City**

This is a real time project where in it includes collecting and analyzing data in a city for a metro rail project before the initiation of the project. This analysis includes which localities are feasible for metro rail? Which route will be optimal for the project? And how well the public will utilize it after completion?

The main problem here is how we can construct the metro in the city with least demolitions and given least compensation to the public. For this project I do not have access to the data as it is sensitive, but from my friend who worked on it, I came to know that they followed clustering techniques and used MFP algorithm for data analysis and finding frequent patterns. With this analysis, they were able to figure out the feasible routes to construct the metro project and in which areas they are more feasible to travel in.

**Online Railway Reservation System**

One more example is analysis of an online railway reservation system. This analysis includes from where requests are received to the server? How many authorized requests are received in peak hour of booking? Response time for each request received. This analysis gave the government to figure out which are authorized reservation centers and which are fake ones. This analysis is done using a K means algorithm where data is clustered based on some criteria.

The main problem here is to identify fake and duplicate reservation centers and do not accept requests from them during peak hours of the time and accept requests from the authorized ones.

I am still continuing my research in this particular domain and figuring out my analysis of data collected in this particular domain. Still need to learn more of machine learning algorithms and data analysis concepts. Currently I am facing a major challenge of data acquisition from data sources. However, I am figuring out how to overcome these challenges.

**References**

1. <http://ieeexplore.ieee.org.ezp1.villanova.edu/stamp/stamp.jsp?tp=&arnumber=5077118>
2. [http://www.neighbourhood.statistics.gov.uk/dissemination/instanceSelection.do?JSAllowed=true&Function=&%24ph=60\_61&CurrentPageId=61&step=2&datasetFamilyId=1201&instanceSelection=119561&Next.x=28&Next.y=16](http://www.neighbourhood.statistics.gov.uk/dissemination/instanceSelection.do?JSAllowed=true&Function=&$ph=60_61&CurrentPageId=61&step=2&datasetFamilyId=1201&instanceSelection=119561&Next.x=28&Next.y=16)
3. <http://docs.scipy.org/doc/scipy/reference/tutorial/spatial.html>
4. <https://data.gov.uk/dataset/land_use_statistics_generalised_land_use_database>
5. [https://www.ordnancesurvey.co.uk/business-and- government/products/topography-layer.html](https://www.ordnancesurvey.co.uk/business-and-%20government/products/topography-layer.html)
6. <http://iveybusinessjournal.com/publication/four-strategies-to-capture-and-create-value-from-big-data/>
7. <http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics>
8. <http://www.webopedia.com/TERM/B/big_data_analytics.html>
9. <http://hanj.cs.illinois.edu/bk2/>
10. <http://ccs1.hnue.edu.vn/hungtd/DM2012/DataMining_BOOK.pdf>
11. <https://msdn.microsoft.com/en-us/library/cc707786(v=sql.105).aspx>
12. P.Thomas, Macredie “Knowledge Discovery and Data Mining”
13. <https://en.wikipedia.org/wiki/Cluster_analysis>
14. <http://www.roselladb.com/geographic-segmentation.htm>
15. Jiawan Han, Micheline Kamber “Data Mining Concepts andTechniques” 2nd edition 2004.
16. <https://data.gov.in/catalogs/sector/Transport-9383#sort_by=created&sort_order=DESC&items_per_page=9&page=1&path=sector/Transport-9383>
17. <https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm>
18. <https://docs.python.org/release/2.5.2/lib/typesmapping.html>