**Project Proposal**

Hector Lopez

**Project Title:**

Analyzing and Adjusting Data Quality on IoT data for Wind Sites

**Abstract**

In the Internet of things (IoT), data gathered from a global-scale deployment of smart-things are the base for making intelligent decisions and providing services. This paper focuses on the IoT data from a Wind Site with many Wind Turbines and aims to implement a method for detecting and tracking data quality (DQ) of the IoT data. The paper aims to identify the definitions of the DQ dimensions specific to the domain for Wind Sites. After generating a method to detect data quality a dataset from a live wind site is used to determine the effectiveness of the DQ strategy. Data capture settings are adjusted to understand the tradeoffs between DQ dimensions. Data is also categorized and analyzed in batches to understand the DQ of the captured dataset. The results explore any improvements or actionable information that can be used to improve the DQ. In conclusion the possibility of future research and enhancements is presented.

**Description:**

1. IoT Data Characteristics
2. Definition of DQ and DQ Dimensions
3. DQ Dimensions for domain-specific application for Smart Grid
4. Finding overlap of DQ Dimensions for Smart Grid and Applying it to Wind Sites
5. Factor affecting DQ in IoT and their impact
   1. Network Bandwidth
   2. Remote Location
   3. Limited Data Storage
   4. Slow Networks
   5. Broken Sensors
   6. Connectivity
   7. Processing Errors
6. DQ Dimensions Trade-offs at Wind Sites
   1. Resolution vs. Completeness
   2. Timeliness vs. Accuracy
7. Present network and logical architecture of a wind site and its collection strategy of IoT data from SCADA and other RTU devices into a Database on the Cloud
8. Design DQ Algorithms for Intrinsic Dimensions
   1. Design algorithms to analyzes DQ in the identified Dimensions given the “Raw” sensor data and the captured data. i.e.:
      1. Resolution : Normalized RMSE between Captured Sensor Data (in the cloud) vs. Raw Sensor Data (from local logs)
      2. Precision : Relative Standard Deviation of Captured Sensor Data vs. Raw Sensor Data
      3. Completeness : Number of tags captured vs. Number of tags on sensor
   2. Using the DQ Analysis algorithms , rewrite them in terms of the hyperparameters and, Identify what the data quality would be if the hyperparameters settings are adjusted in the capturing stage.
      1. If capturing at lower frequency than sensor data
      2. If capturing less data than sensor data
      3. If data is filtered based on deadbands
9. Design DQ Algorithms for Contextual Dimensions
   1. Understand the contextual features of the data
      1. Identify cross-corelated data (i.e. Reactive Power, Active Power, Frequency)
      2. Identify continuous data (temperature, windspeed, etc.)
      3. Identify smooth variation data (humidity, pressure , etc.)
      4. Identify periodic data
   2. Design algorithms to take advantage of the domain specific data context and provide outlier detection to improve data quality analysis
      1. Cross-correlated data should have a correlation
      2. Continuous, smooth variation data should not have jumps
         1. Detecting sharp slopes inside small windows
      3. Periodic data should be able to show periodicity in a given amount of time
         1. Using sampling methods for periodicity of high frequency
         2. Attempt to employ Regression methods to identify periodicity outside of the given dataset (for large periodic data)
10. Given a capture device setting and a batch of data from IoT Data set analyze the data using the DQ Intrinsic and Contextual Dimension algorithms and determine the overall quality of the data as well as granular quality of the sensor data
    1. A data capturing device “Data Logger“, has settings that can effect the DQ
    2. Given the settings determine the effect it has to the sites DQ when capturing the data
    3. With a large batch of data from the site apply the algorithms to the large dataset to determine its data quality. Does the overall data quality change based on the size of the batch? How much data do I need to stabilize the overall data quality measurement.
11. Propose methods to adjust trade-offs
    1. By how much do I need to adjust the settings in the Data Logger to increase the Data Quality?

**References :**

1. Aimad Karkouch “Data Quality in Internet of Things: A state of the art servey” 2016

2. [Sabrina Sicaria](https://www.sciencedirect.com/science/article/pii/S0306437916300072" \l "!), [Alessandra R](https://www.sciencedirect.com/science/article/pii/S0306437916300072#!) “A secure and quality-aware prototypical architecture for the Internet of Things” 2016

2. Zita Vale,Hugo Morais,Pedro Faria,Carlos Ramos Distribution system operation supported by contextual energy resource management

based on intelligent SCADA , GECAD – Knowledge Engineering and Decision Support

Research Center, Polytechnic of Porto (IPP), R. Dr. António Bernardino de Almeida, 431, 4200-072 Porto, Portugal

3. K.Metaxiotis A.Kagiannas D.Askounis J.Psarras , Artificial intelligence in short term

electric load forecasting: a state-of-the-art survey for the researcher

Department of Electrical and Computer Engineering, National Technical University

Athens, 9 Iroon Polytechniou str., Zografou 15773, Athens, Greece

Received 12 February 2002, Accepted 10 June 2002, Available online 15 October 2002.