OSWEGO

STATE UNIVERSITY OF NEW YORK

SUNY Oswego Information Science:

Generalized Overview

Fall 2017

ISC 496

Author: Joshua Harkness

Date: 10/8/2017

Style of Citation: APA

While many people are just learning about the vast field of information science and its relationship with computers and technology, information science has been around as long as humans have been able to communicate. The modern-day view of the field cannot be ignored though. With the continued integration of computers in everyday life, a surplus of data now exists and we need to find uses for it. The collection process, as well as the manipulation of the data, play vital roles which allow us to make use of the data through analysis.

While a significant amount of data exists, it would be useless without being able to accurately and appropriately collect it. “Accurate data collection is essential to maintaining the integrity of research.” (Responsible). Unfortunately, throughout my academic career, I did not gain much experience in the collection of data. Fortunately, I was able to gain some collection experience through an internship with Prof. Isabelle Bichindaritz.

In summary, the internship with Dr. Bichindaritz strived to discover which genes in a known gene set were more likely to lead to breast cancer relapse. One of the first questions posed in the project was “where do we get our data from?” Whether the data involves possibly saving lives through cancer research or determining where a pitcher may throw the next pitch, the data must be reliable. The genome data used for the project came directly from the National Cancer Institute. Using the data from such an established institute helps remove ambiguity. If the data for the genes came from a neurosurgeon instead, the data would either be significantly less relevant, or be much less reliable due to the different fields of study. Once collected, data is rarely ready for immediate analysis.

The information science program constantly emphasized the importance of data manipulation. Usable data is often stored in tables designed to accentuate one specific variable. Frequently, relationships between the tables must occur for more advanced analysis. In my freshman class, we were given data to organize into tables and then asked to create relationships between the tables. The data dealt with information regarding a reality agency. Each table alone only provided information on individual objects; the object could be a tenant, branch of the business, employee, etc. Without forming relationships between the tables, analysis would be significantly limited in what it could discover. Manipulating the data by transferring it to relationships or performing algebraic expressions to create new data allows for much more complex analysis. One example of a more complex analysis would be calculating the rent owed by a tenant. A rent paid variable could be compared against a rent required variable for a specific apartment. The difference could be calculated and stored into a new variable, called rent owed. More direct examples of manipulation occur too.

Occasionally, tables must be merged and the variables within require complete manipulation. Another example derives from the internship. The main portion of my work involved the manipulation of data in one table in order for it to merge properly with another. Without manipulating the data, the program I was using wouldn't have recognized which variables were related. The variable names were slightly different between the two tables. In order to correct the difference, I used string manipulation on the proper column names. By manipulating the column names, we were able to introduce 3719 new variables to the table. This addition allowed for a more accurate analysis.

Analytics is arguably the most important aspect of information science. All the other subfields lead to the goal of performing safe, precise analytics. In order understand which direction to move towards in data analytics, the analyzer must have a strong statistical background.

The statistics classes taken at the undergraduate level help build the statistics rapport. The classes helped build an understanding in correlation, sampling, regression and hypothesis testing. One of the statistical tests I ran was an attempt to find an edge over my friends in a videogame. Unfortunately, after collecting my data and manipulating it for analysis, I could not find any strong correlations to use to my advantage. The gaming example only used a narrow set of data. For a more complete analysis attempt, animal classification was used.

In an upper level course, multiple statistical analyses were used as an attempt to classify animals based on their characteristics similar to how the animal kingdom already categorizes them. The first attempt used was to determine K using the Ward algorithm. Secondly, the K means algorithm was used again, but instead of relying on the Ward algorithm to determine K, it was given 7; the ideal matching number. Lastly, in hopes of determining a more accurate result, Kohonen self-organizing maps were used. The K means algorithm ended up being the more accurate result. While these algorithms sound impressive, I must admit I do not know what they are and that I was directed to use them. Without knowing the algorithms, there is no feasible way the proper algorithm would be discovered for the data provided.

In time, I will build a stronger statistical background to further my understanding of data analytics, but the foundation was needed first. Without understanding the collection process there would be no clean, accurate data for me to analyze. Understanding how to manipulate the data for analysis builds a stronger understanding of how the data needs to be analyzed. A person cannot manipulate the data unless they understand what they want it to become and why that particular data needs to change for the analysis. Analysis may be the most important aspect of information science, but the others play pivotal roles which must be understood first. A house is only as reliable as the foundation it's built upon.

Reference

Responsible Conduct in Data Management. (n.d.). Retrieved October 08, 2017, from https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dctopic.html