Peter Mattoon

SBIR Grant

Overview, Key Words, and Subtopic Name:

A website is being created to analyze sets of data in order to draw conclusions from that data. The website will start by allowing the user to input one or more sets of data and then ask for the type of analysis desired. After analyzing the data the website will return its findings along with a graphical representation of the data in order to make it more easily understood.

Key words: Analysis, Big Data, Statistical Analysis, Graph, Website, Computer Science, Software

Subtopic: Information, Computer Science and Engineering (EA5)

Intellectual Merit:

This Small Business Innovation Research Phase I Project will be designing an algorithm to make multiple different analyses on a data set in a reasonable amount of time while still providing an accurate and comprehensive analysis of the entire data set. The creation of such an algorithm is technologically difficult because the software must be programed to sample enough data for an accurate analysis without wasting time sampling extraneous data. The algorithm must also be able to determine which data sets may be analyzed together, such as ones that were collected over time versus ones that cannot be compared against each other. The goal of this research project will be achieved through rigorous experimentation on a sample data set and when completed it will provide an abundance of tools to analyze data.

Broader/Commercial Impact:

The research project described above will be extremely important to smaller businesses in the coming years because of the increasing reliance of companies on big data. All businesses big and small collect huge amounts of data on their customers. Large companies are starting to utilize that data to better understand the needs and wants of their customers so that they may offer a better experience in the future. The currently large cost and time consuming nature of such data analysis means that small businesses are losing to their competition; however this online analysis tool will be able to help even the field at an affordable price and in a timely manner. Even for those without a commercial need will be able to use the website to answer questions and understand data on just about subject.

Elevator Pitch:

The website created from this project will be and do a host of different things for a wide variety of people in a fast, easy and understandable way. The website will be able to analyze large sets of data supplied by the user and will return an understandable analysis. However, while the website will only be able to analyzes data, the kind of data and analysis will differ from one user to another. The main focus of the website will be analysis for smaller businesses who cannot afford to hire someone but truly anyone will be able to use the website. Since it will easy to use and will return an easily understandable answer.

The value that the website will bring to its customers relies upon its ease of use, its affordability and its meaningful analysis. The website will be easy to use because it will be a two-step process for any users. First they will have to upload their data in a clear format for the website to then read into a database. Then they need select the type of analysis they would like done on their data and that is it. All they need to do after that is wait for the data to be processed. Furthermore, the website will either be free with ads or have a small membership fee depending upon serve costs to run the website. Lastly, and most importantly the website will be able to supply the user with an easily understood and meaningful analysis of their data, which will include graphs that fit the type of analysis carried out on the user's data. The website will be a product that anyone will feel comfortable using and easily understand.

Moreover, while the website can be used by anyone with a dataset, the target customers are small businesses looking to improve the experience of their customers. The small businesses are the target consumer of the website because they have the data and a need to analyze and understand it at a low cost. Small businesses collect huge amounts of data on their customers in the normal course of business but mostly they do not do anything with that data either because of a lack of time, understanding or both. However, they all want to because they know it will lead to an improved customer experience. They cannot outsource the analysis because of cost and some just do not understand how to get anything meaningful from the data. That is why the website will be a perfect fit for small businesses trying to improve their customer experience and compete with larger businesses that can afford to pay for data analysis.

The website is not going to be the industry leader in data analysis but it is going to be extremely easy to use and understand. The data will be supplied by the user and the analysis selected and then after running it through our software the website will return the data in a meaningful and comprehensible format. That is the area in which the website is going to innovate.

In conclusion the website will be an affordable and valuable tool for anyone, but especially small businesses to turn their data into knowledge about their customers.

Commercial Opportunity:

The data analysis website will be competing, for most of its business, in the lower end of the consulting market. A growing market that as of 2013 had a value of around 40 billion dollars, however we are not going after the entire market but instead we are targeting the lower and less well served end. The ideal client for the our website is a small business looking to improve its ability to service its customers. As every business becomes more reliant on technology for every part of their activities they inevitably collect data on their own customers, as well as their suppliers and simply how the business operates. All of this data can be extremely useful to a company but it has to be analysed and properly understood before it can become anything more than a large amount of confusing numbers and letters. Large businesses can afford to higher consulting firms that will analysis everything and then provide recommendations on how to improve but smaller firms and businesses do not have the capital to afford a data analyst that can charge up to 250 dollars per hour. The consulting and data analytics marketplace is full of high end firms but the lower end is very underserved because a profit cannot be made with the current model.

The key risk involved in the data analysis website is that it cannot be protected from others who wish to copy it. The website is built to provide analysis of data but the way which the data is analysed is not unique. The platform and ease of use are new, however, a common statistical algorithm cannot be patent protected and new statistical algorithms cannot be invented because that would lower the credibility of the website. That is why to be successful this website must be the first on the market in order that we secure a large market share and loyal customers. After we first draw in our customers we can insure they stay with the site by consistently increasing, the forms of analysis, ease of use and the clarity of the results.

However, while having a large market share is important for any business it is useless without any way to turn those customers into profit and the way that the website will do that is through ads and memberships. Ads for almost any website are an obvious choice but if the right ads are found a significant profit can still be made. Many websites today rely on google's ad service to get ads for their websites and while this works it is not the best option. First, because there may not be any synergy with our website and they are not the best paying. The best paying ads is when another company directly pays for space on a website and we will attract that type of ad by reaching out to other small business focused companies. The second source of revenue will be through memberships, which will provide extra services to our clients. Our clients will need those extra services to analyze any truly large data sets because the free version will have file size constraints, due to time concerns. Furthermore, customers will also be able to access more forms of analysis with the memberships that are too computationally intensive to run for free. Overall, our website is positioned to be a strong business in the future.

Societal Impact:

One of the goals of this project is to provide a way for small businesses that so far have been left behind by the increasing importance to big data to start to catch up to their larger competition. This product will allow the smaller players to continue to continue to compete in today's economy and in doing so help the economy as a whole since competition is the cornerstone upon which capitalism is built. If used wisely the data analytics website will allow small business to better understand how to serve their own customers with an easier and more enjoyable experience. And so if the product is widely used consumers everywhere will have more enjoyable experience throughout their normal lives.

Thankfully the data analysis website is very safe and will not do any harm if used correctly. Environmentally, other than normal electricity costs the website will not have any negative effects, furthermore, the website could lead others into being more efficient with their use of resources and in doing so help reduce the need for natural resources. When it comes to health effects the changes will only be positive due to increased customer experiences at health care providers. Moreover, in the case of children, if they use the website at all, it will only leave them more knowledgeable.

The only worrying aspect of the website is that it can be used by anyone with any data and so it is accessible to criminals as well. Fortunately, the system would require data sets from the criminals to be of any use and that requirement would later help authorities to convict them when they are caught. Overall, the website will do a lot of good for the world and its customers.

Technical Discussion and R&D Plan:

The key goal of this research and development project is to allow anyone who wishes to be able to upload data to our website and in return receive an output that is both understandable and meaningful. To reach that goal requires many innovations but a few of the most important are the creation of a graphing algorithm that will decide on the best graph to display the input data and a algorithm that decides how to best annotate the graph once it is created.

The algorithm that will decide on the correct graph will be run by a strong AI using rules that I have created. As of now the algorithm will be dealing with four different graphs: bar, pie, scatter and line. Each of these types of graph are best at showing certain types of data and so to best make data understandable it is important that the correct graph is used. For example if one wanted to examine the change in population within each county of Massachusetts between 2010 and 2015 a bar graph would be perfect since it would be able to show the change between the two dates and make clear that each county is different. A pie chart would however, be the best choice if one only wanted to look at only one year and see how much of the total massachusetts each county represents. A scatter plot would work well if one wanted to look at how the size of each county physically relates to how large the population is. Lastly, a line graph would excel at showing the change in a few counties over a long period of time with many points.

Every type of graph excels in different areas and each graph implicitly and explicitly shows data to a user. Bar graphs are best at showing large changes between small to medium size sets of data. Pie charts on the other hand are great at presenting small data sets that are all part of a single whole. Scatter plots have a large range of different uses but are best at examining two related datasets and showing how they relate and if there are any clusters or outliers. Finally, line graphs can handle large data sets but work best when that data is collected over time since the very fact that all of the points on the graph are connected implies that one point follows from another.

The other innovation of this project is that after the correct graph is created for each data set those graphs will then be annotated to best make their findings as clear as possible for the user. Some of the annotations are obvious such as finding the maximum and minimum on a bar or line graph or finding outliers on a scatter plot. Others, such as finding the skew of the data are more interesting however, what makes this innovative is the fact that the annotations will be automatically generated in the best possible way to make the data clear to the user what it means.

Identifying which graph to use for a given set of data while be technically difficult to accomplish for a few different reasons. First, the factors within the data must be identified that show which graph would be best. Second, that data must then be interpreted by the program in order to determine which graph is best. For some data sets this will be easy since it might heavily lean towards one type of graph however, others may have conflicting signals and so different factors will need to be weighted against each other to determine the most important. The best way to do make these decisions is to create a strong AI, meaning a machine learning algorithm that is based off of a set of rules but also using training data to come up with a solution.

However, the creation of a machine learning algorithm comes with its own problems. Even after the factors in the data have been found the creation of rules is not easy since some factors may point to a few different graphs or work in relation with other factors and so in some situations mean one thing but in others a totally different graph should be chosen. Furthermore, even after the rules have been written a training set of data needs to be made that will train the program on the best way to interpret different factors when the conflicting answers are given.

In the case of annotating the graph after it has been created the technical difficulties lay more in determining how and when an annotation should be created and displayed then in the annotation of the graph itself. This is because every data set can be analyzed in a multitude of ways however if all of those were to be displayed it would overwhelm the user and in doing so complete defeat the purpose of having those annotations in the first place. They are meant to inform the user not scare them away. An algorithm will need to be built to deal with this problem as well but in this case it will rank the different analyses on how important each one is and if possible how informative it will be to a user all while making sure everything stays simple enough that the average user is not scared away by overwhelming or esoteric data.

Key Objectives

There are a few different key objectives to complete this phase of the research, the first of which is the creation of an algorithm to determine the best way to graph a given set of data. To create such an algorithm the first thing that needs to be identified is what characteristics about a data set determine the best way to graph that data. After that rules must be written around those characteristics and then training data must be created in order to make the algorithm using machine learning.

Another, key area of research is the algorithm to correctly annotate the graphs after they have been created and to do that a few questions need to answered. What is most important for a person to know about data in a graph? How many annotations are enough and how many are too many? How complicated should the annotations be, meaning should it stop at the level of trend lines or should the skew of data be included? Where on the graph should annotations be located? Should a user be given options or should the entire process be decided by the algorithm without user input?

After both algorithms have been created in order that the program be accessible to the general public a website will need to be created and then integrated with the rest of the code. The website will need to be able to upload files to the program and then analyse and visualize that data to the specifications of the user.

The last big milestone will be determining the best way to profit from the creation of this website. Is the best way to have ads on the website or will that just annoy people? Are there features that can be offered that would make some people willing to purchase a membership? Are people willing to pay to be able to work with larger sets of data that take more processing power than the average request. Overall, there are still many milestones that need to be reached before this project comes to fruition and becomes commercially feasible and profitable.

Technical Milestones

There are a few technical milestones that must be met before the this project can be brought to market, the first of which is the creation of the graphing algorithm to determine which graph is best suited for a set of data. This algorithm being one of the selling points of this product until it is completed there is no possible way this project could be considered finished. The second technical accomplishment needed for completion is the creation of the annotation algorithm another fundamental part of this project. Lastly, after the creation of the algorithms a website must be created to host the program and interface with users and after that website is completed it must be integrated with all of the rest of the code that makes up this project.

Timeline

- September
    - Decide on the focus of the project and research past projects in this field
- October
    - Begin creation of project
- November
    - Continue to write code for project focusing the most basic features needed to build later innovations
    - Present on the overall plan of the project
- Early December
    - Demonstrate to professor Simha
- Mid December
    - The creation of a rule based version of the graph and annotation algorithms
        - This will be presented to professor Wood
- Mid January
    - A more complete version of both of the algorithms
        - This will be accomplished with many experiments to find the most important factors and how to weight them along with the training data created to teach a machine learning algorithm
        - This will be displayed to professor Wood
- Late January
    - The creation of a basic website
- Mid February
    - Integration of all code
- Till the end
    - Continued refinement of all of the algorithms and features
    - Expansion of current features
    - Addition of new features as time permits
- Final Project Presentation

Revisions

After the peer revision section I expanded on the key objective section that had previously been made up of bullet points. I also made the timeline section more clear after comments were made that it was confusingly written.