Eric Downing

Bryan Ford

11/4/2016

Initial Setup

Installed

Python

R

Github

Cygwin

SSH into [eric@68.36.180.156](mailto:eric@68.36.180.156)

Technologies we plan to use

Hadoop

RStudio

Ggplot2

First Steps

1. Began by focusing and learning about the aspect of what we want to do in regards to the log data that we have.
2. Read paper: <https://users.cs.fiu.edu/~taoli/pub/Li-sigkddExp.pdf> to get a better understanding of the challenge presented before us.
3. After reading the article, there are a few relevant things that could be deduced from understanding all log information.
   1. With a good classification of each of the log items, if one were to properly categorize them, they could all be grouped and introduced into different states
   2. With each of them into a new state, we could begin to get probabilities of the occurrence of future states given the present state.
   3. Once all this has been accomplished we can then narrow it down to probabilities of generating an error state.
      1. From here we could then begin to reduce that probability by examining the current state, or the one that would lead to an error.
4. Different Algorithms that are useful:
   1. Naïve Bayes
   2. Forward
   3. Baum-Welch
   4. Viterbi
5. (Hidden) Markov Models
   1. Markov Chain (being the simplest) would give probabilities of a transition to the next states.
6. Mapping Certain Keywords together could help in creating a more generalized system
   1. In the article, it was explained that between Linux and Windows machines there was a different context that necessarily didn’t mean different things but would add difficulty to the categorization process.
      1. It may not be a bad idea to have a map of certain keywords so that categorization is improved.
7. Given the temporal information, we could come up with a way to deduce the possibilities of errors in a given time frame and may be able to correlate that to a different event.
8. Possibility of writing a parser that would go through a log file and insert all relevant information into a database
   1. Here we could use our categorizer and implement it this way
   2. Having the info in a data store would allow us to access information faster than would reading it every time from a text file.
      1. One and done
   3. With the ability to create other tables and what not from the store we could build some bridging tables and further develop our understanding of the data being given from the logs
   4. If written correctly, this could be a general program which could be used for any other necessary items in the future
   5. Could make a WPF application that would allow us to do as such
9. Possible Visual representations of data
   1. This is not necessary, however, being able to present info is never bad
   2. Given all the information and displaying everything with its given timestamp we could potentially see a point that errors occur and make preparations for when they are about to occur
10. Warehouse Management
    1. Since we are stowing all of our items in containers, data mining might be able to accurately predict what orders are going to be placed at a specific time. Going off of probabilities we may be able to indicate which product is going faster than others.
       1. Storing items in closer position relative to the crane essentially equals out to an increase in efficiency. While this may not be practical in storing all items up front, dynamically being able to decide how many items of a general variety up front may turn into an increase in efficient storage and increase through put.
11. While we’re able to have Hadoop spread a data query across a cluster the only thing that would limit us on this end is resources.
12. Emulations of how a crane would perform given general input and knowledge of where to put items.
    1. This could be a test to see how any given crane would perform given an accurate account of data from WFS
    2. Comparison between our information that we have now and the predictability information
    3. Hooking up this emulation to a database