Analysis of Operator Mailing Lists with Real-time Network Measurements

Guanyu Zhu, Wei-Ting Lin, Zhaowei Sun

Motivation & Related Work

Internet outages are an essential topic for the contemporary society because of the popularity of mobile devices rises, and the broad scope existence of Internet services. A sudden Internet outages could cause several consequences such as companies are unable to work [1], students are unable to do their assignments [2] and even the finance of a country could drop down in a short period of time. If there is a way that can help us to analyze and predict the causes of Internet outages, Internet providers and the technicians will be able to solve the problems and repair the hardware more efficiently. Unfortunately, although people have already noticed how critical it is, the study of Internet outages is being obstructed by many reasons such as the benefits of the Internet providers, private information, and the inadequate open resources. One related paper [3] puts great effort on the Internet outage this topic, the authors use Natural Language Processing (NLP) and Machine Learning technique to analyze and categorize the keywords in the outage mailing list [4] in order to classify the cause and effect of the Internet outages. The purpose of this project is to follow the paper [3] and spur real time measurements of the incident in the mailing list to improve the network related works in order to move the Internet outage this topic to a further stage.

Plan & Method:

We probably divide our project into 3 steps:
1.We should do some data preprocessing works to extract useful information from our dataset -- the outages mailing list. In this step, we can use the natural language processing and text mining techniques. First we can focus on the Thread level of datasets and then remove some spurious data and stop-words using some available tools like the Stanford CoreNLP toolkit [3]. After finish this step, we should get many key words about the outage information.

2. We should use some machine learning methods to classify the type of outages based on the information that is extracted in step 1. First if we want to use the supervised learning method to do the classification, we need to label some training data based on the domain knowledge. Besides, we can use the Fleiss' k metric to make the label consistent [3]. Then we can use any machine learning method like SVM or Naive Bayes or Decision Tree to construct the classifier using the labeling training data. Finally we can use our constructed classifier to classify the types of outages in the outage mailing list. After this step, we should get a methodology to characterizing the causes of failures.

3. We should use the technique presented in step 1 and 2 to implement the real time Internet measurement.
We can focus on some aspects of the Network like the traceroutes, DDOS attacks, some performance of the network and so on. We think this is a difference with other research.

Expectation:

Based on the dataset mentioned above, we want to get the reliability issues spanning multiple networks over long time using the text mining, NLP, and machine learning techniques, such as the categories of outage cause and type of outage. When the incidents occur in the mailing list, we can firstly classify the type of this outage based the previous work and spur real-time diagnostics about the network measurements. For example, when we get the outage and figure out this outage is caused by pack loss, then we can use RIPE Atlas to get the info of traceroute and so on.

Timeline:

Feb.16 -- Apr.1
-- Data preprocessing
-- Feature extraction using NLP techniques
-- Classification the types of outages using machine methods.
Apr.1 -- Apr.5
-- Discuss the second part of project
-- Discuss several questions about second part of project:
1. Is outage mail sent in real-time by someone?

2. If the outage is recovered right now, so what should I do for the diagnostic in this situation?
3. What's the meaning of real-time?

-- Make detail plan for the real-time network measurements implementation and diagnostics.
-- Prepare the Midterm Report
Apr.6
-- Midterm Report
Apr.1 -- May.10
-- Implement the real-time network measurements system
-- Using some tools do the diagnostics
May.10
-- Final Report

References:

[1] Kristen Carosa (2014, Dec 11) *Widespread FairPoint Internet outage affects NH customers* Retrieved from http://www.wmur.com/money/widespread-fairpoint-internet-outage-affecting-nh-customers/30176172

[2] Mary Scott (2014, September 5) *Pellissippi State internet outage impacts all 5 campuses.* Retrieved from http://www.wbir.com/story/news/local/2014/09/05/pellissippi-state-internet-outage-impacts-all-5-campuses/15152481/

[3] Ritwik Banerjee, Abbas Razaghpanah, Luis Chiang, Akassh Mishra, Vyas Sekar, Yejin Choi, Phillipa Gill *Internet Outages, the Eyewitness Accounts: Analysis of the Outages Mailing List*, 2013

[4] V.Rode. Outages–outages(planned &unplanned) reporting.
https://puck.nether.net/mailman/listinfo/outages