Network Outages Analysis and Real-Time Prediction

Guanyu Zhu
zhuguanyu2010@gmail.com

Wei-Ting Lin
wei-ting.lin@stonybrook.edu

Zhaowei Sun
zhaowei.sun@stonybrook.edu

ABSTRACT

Internet outages are an essential topic for the contemporary society because of the popularity of mobile devices are on the rise, and the broad scope existence of Internet services. A single Internet outage could cause serious impact such as companies are unable to work [1], students are unable to do their assignments[2] and even the finance of a country could be dropped down. As a result, knowing the reasons why an Internet outage happens is desirable. 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 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 goal of this project is that bring such concepts from the paper, increase the number of data-set, use a different way of labeling method and to predict the on-going Internet outage. Imagine if we can analyze and predict the possible causes of an on-going Internet outage, Internet Service Providers and Internet maintenance staff can take this information into account and increase the efficiency while they are repairing an on-going Internet outage.

1. DATASETS

In this section, we introduce the basic idea about the data that we use in this project such as where the data is obtained, what the data is use for, what the data looks like and how do we use the data for our project.

We obtain our data from the outages mailing list[7], which is basically a platform for both network operators and end users to post and discuss the outages that are relevant to the failures of communication infrastructure component. The list contains outage reports, afterwards analysis and discussions on troubleshooting.

We download and analyze the outages mailing list taken on March, 2015 containing threads since its inception on Sep, 2006 [7]. It contains ten years discussions on the outages mailing list. These discussions are organized into thousands of threads. Each thread contains a host-post, and it might also contain none to several replies. However, no matter a host-post or a reply, each of them contains contributor's information, subject, message, system formatted information and an unique message ID. In our implementation, the usage of this data is to extract the subject and the contain of each host-post or reply, and we assign each contain with the same subject to the same thread. In Figure 1, we show the first email, last email, total amount of posts, replies, threads and contributors.

First Email	Sep, 29 2006
Last Email	March, 24 2015
Num of Posts	6963
Num of Replies	2102
Num of Threads	4725
Num of Posters	1256

Figure 1: Datasets

Apparently, the number of replies is always lower than the number of posts. The reason why the number of threads is lower than the combination between the number of posts and the number of replies is because there are some unsubscribed emails which should not be counted into our implementation. Even though the number of posts and the number of replies are fluctuated in this ten years, Figure 2 shows that the number of threads is very evenly distributed in each month.

We also analyze our data-set from four different angles such as Content providers, Internet Service providers, Protocols and Security, and we show them in the Figure 3. From the Content provider graph, we see large amount of words are related to Google on 2014. We searched on the internet and found the fact that Google down several times on Jan, Jun, July 2014. The users couldn't access Gmail, Google+ and even Google colander at the time. Next, we searched the what happened to Verizon on 2009. It shows that the cellular issue were getting attention which matches the fact that the first iPhone were released on July, 2008.

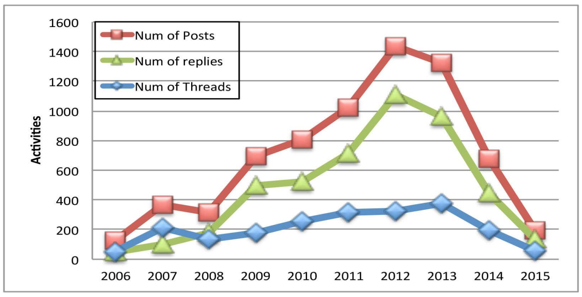


Figure 2: The number of the threads is evenly distributed compares with the number of posts and replies from 2006 - 2015

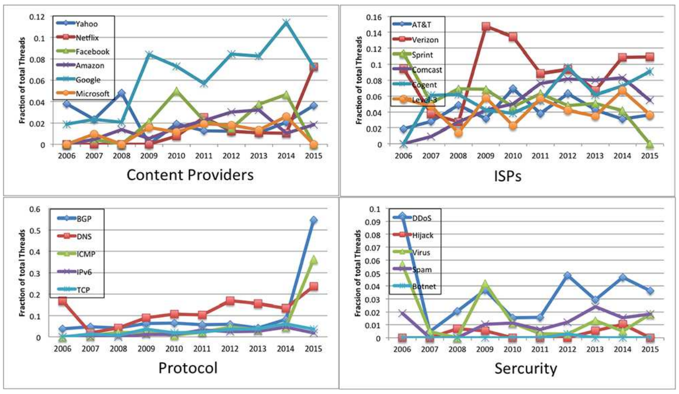


Figure 3: Analyze the dataset from four different perspective.

2. DATA PROCESSING

In this section, we discuss how we extract useful information and how we omit the unimportant information from the outages mailing list.

The outages mailing list are text based messages, which means that it has rich semantic information underlying the Internet outages. Hence, it also presents a challenge in terms of automatically parsing and processing the data. To address this challenge we employ techniques from text mining and natural language processing (NLP).

2.1 Merge the posts that belongs to the same threads

In general, we consider the dataset at the level of threads. Each thread consists of the set of e-mail messages (posts) in the thread. For each thread we extract relevant information (e.g: term and phrases). After removing quoted text (text from previous emails in the thread included in each email) from its posts, we remove the content which is unimportant and repeating such as the content between *"BEGIN PGP SIGNATURE"* and *"END PGP SIGNATURE"*, empty lines, contributors information (e.g: name) and post information (ex: date).

2.2 Remove words that are not related to network outages

In this part, we remove the irrelevant words. In term of "irrelevant words", we mean the words that are not useless for analyzing the network outages. We classify those irrelevant words into 9 categories and show them below.

1. Spurious data. We remove spurious data, which contained the identifying e-mail signatures used by posters and some data added by system or antivirus software. For example, "This message has been scanned for viruses and dangerous content by MailScanner, and is believed to be clean." We treated this kind of messages as the spurious data and should be discarded.

2. Links. Then we ignored the url, website links and email links in the posts. Those are has little things with the outage of network.

3. Punctuations and Numbers.

4. Traceroute measurements. We think these info are useless because only based on the traceroute measurements we can't figure out the root cause of an incident.

5. Stop words(e.g., articles, prepositions and pronouns). We also use a list of stop words obtained from the SMART information retrieval system[5].

6. Organization and Human names. These organization and Human names are no meaning for us to analyze the cause of outage, such as Sprint, AT&T, Gary, Tim, etc.

7. Time-related and Place-related words. Such as day, night, NYC, San Jose, etc.

8. Some unrelated abbreviation words. Such like ICS, ISP, etc.

9. Others. This includes some entities words(like issue,information, etc) or phrase(like "in order to") that have nothing with network but can affect the efficiency and accuracy about the NLP(natural language processing) analysis. . .

Compared with the methods mentioned in the reference[3] (which only removes about 4 kinds of words in the above list), removing words that are listed in the list above makes our NLP analysis be more accurate and efficient.

2.3 Stemming and Lemmatization

After step 2, the remaining words should be stemmed and lemmatized (grouping the different inflected forms of a word) using python Natural Language Toolkit(NLTK) so they can be analyzed as a single item. For example, determining that "walk", "walked" and "walking" are all forms of the same verb: "to walk". Note that the simple stemming (i.e., walking to walk) does not suffice as it cannot differentiate the parts of speech based on context: e.g., when the term "meeting" acts as a verb: "we are meeting tomorrow" vs. a noun "let's go to the meeting".

The reason for doing stemming and Lemmatization is to decrease the dimension of the data, because person and persons have the same effect and meaning in the data for classifying the outage type, if we regard them as different word, it does not improve the classification effect but increase the dimension of the data, it will decrease the efficiency of running time and even the accuracy of our classification.

2.4 TF-IDF

After the step 3, at first, our initial idea is to use TF-IDF algorithm[6] of python Natural Language Toolkit (NLTK) to filter out words with tf-idf values less than 0.2. The reason why we choose 0.2 is not only because the words with low tf-idf value indicates that the words are very common throughout the dataset, and it is also because those words are not useful for the classification method to classify the type of outage. But after we use TF-IDF algorithm to get the high value words, we found that some high tf-idf value words also have no effect for our outage type classification.

2.5 Generate 2-dimension matrix for classification

After step 4 we recompute the term frequency for each word in the dataset and generate the 2-dimension matrix to store these term frequency. Every row indicates the different thread, every column indicates the different words that appear in the dataset. Once we get the matrix, we can use this matrix to do the classification because this is the "true" data that we want. The threshold was chosen such that it filtered out the bottom 29% of terms in terms of tf-idf value

3. CLASSIFICATION METHODOLOGY

The terms and phrases extracted in our initial processing give a high-level view of the discussions on the mailing list. In this section, we discuss a classification methodology to help us systematically categorize the outages over time.

3.1 Labeling

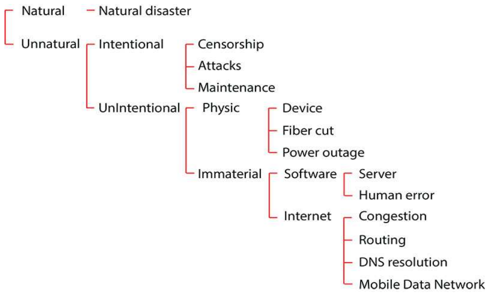


Figure 4: Labeling Criterion

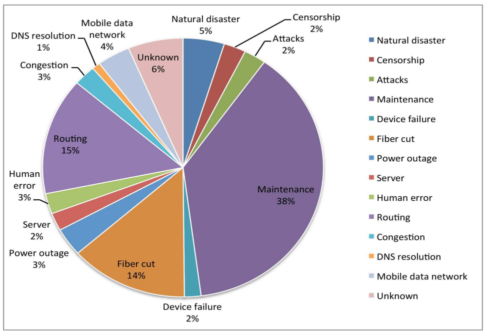


Figure 5: Outage types distribution in 315 threads

First, based on our network knowledge and general network outage types, we classify outages into 13 different types: Routing, Power Outage, Natural Disaster, Mobile Data Network, Fiber Cut, DNS Resolution, Device Failure, Congestion, Censorship, Attack, Maintenance, Server and Human error[3]. In addition, beside these 13 outage types, we add one more unknown category because there are some messages having inadequate information to define which outage types that they should be. Figure 4 shows how we categorize Internet outage types from a big scope to specific.[1]

Our goal is to automatically characterize each outage e-mail thread into categories along these dimensions. However, because computers do not have the network knowledge, sometimes labeling task runs into ambiguity. For example, an earthquake damages the cables in a region and the damage cables cause the Internet outage. Should this outage be classified into Natural disaster or fiber cut? Even for human this answer is ambiguous, no need to mention how difficult it would be for a computer without network knowledge base. Hence, we define an addition category "unknown" to deal with the ambiguity. Next, because of the huge amount of data in the dataset, each of us label 315 threads (107 for training, 108 for testing accuracy). The outage type distribution is shown in Figure 5. To validate that our manual annotations are consistent, we use the Fleiss' metric [8]; the

[1] Note: we do not indicate that this labeling criterion is absolutely right. We define this labeling criterion based on our knowledge and the observed outage types in the dataset.

value was 0.63 for the outage types, which is considered a ”Substantial agreement” [8]. Given this confidence, we use these manual labels to bootstrap our learning process described below.

3.2 Choice of algorithm

In the previous part, we know that the outage type is discrete, so we can use document classification method to solve the problem. Because our dataset is a large number of mail texts that includes many distinct words, we can analyze our datasets as a bag-of word problem. Bag-of word model is a simplifying representation used in natural language processing and information retrieval. In this model, a text is represented as the multiset of its words, disregarding grammar and word order but keeping multiplicity. Bag-of-word model is often used in document classification problem, where the occurrence of each word is used as a feature for training a classifier.[9] So it is better using supervised learning than the unsupervised learning. However, due to the dataset is huge, manually labeling all data is meaningless. Hence, we decide to use semi-supervised learning, we label a small part of data (about 15% of our dataset), use the remaining large amount of unlabeled data (about 85%) to help training the labeled data. We found that this method fits our dataset well and produces considerable improvement in learning accuracy.

3.2.1 Training the dataset

If we use multi-classification method, the difficulty will increase largely because the types are multiple (14 types), which means that it is low efficient, time consuming and worse result. So we simplify it to be a multiple binary classification problem. Instead of partitioning the dataset into 14 categories, we determine whether a thread belongs in a particular category or not. So based on this method, we should classify the dataset 14 times to get all type of outage classification. Compared to classify the dataset one time using multi-classification, this method largely decreases the difficulty and improve the efficiency. For solving the binary classification problem, we use semi-supervised SVM model. We halve labeled 315 threads. 107 labeled data plus the unlabeled data as our training data. And, 108 for the test data which are used in next part. In the supervised learning part, SVM is a good classifier, in the unsupervised learning part, we use EM algorithm to use unlabeled data help the supervised learning.

3.2.2 Testing and Evaluation

After training, we use the test data to evaluate our classifier. We get the accuracy of every binary classifier and the average accuracy of all 14 binary classifiers. The Figure 6 shows these accuracy. From the figure we can see that the average accuracy is almost 80%. We think this accuracy achieves our purpose. The variance between every binary classifier's accuracy and the average accuracy is only 0.29%. This indicates that our classifier can handle a general case but not extreme cases. However, it is enough for us to predict the new and unlabeled data in general.

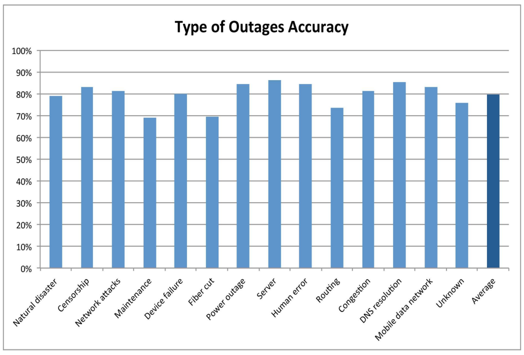


Figure 6: Classification accuracy

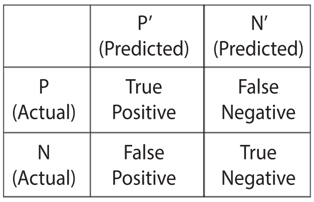
3.2.3 Prediction of the unlabeled data

In this part, we want to predict all our unlabeled data because we think the accuracy of the classification is good. But after the prediction, we found that some threads belong to many categories and some threads don't belong to any category. We analyze that this is a tradeoff of simplifying the classification method. At first we decided to assign the outage types which has the highest accuracy to the threads which belong to many categories and assign the “unknown” types to the threads which don't belong to any category. However, this method did not entirely solve the problem. Thus, we decide to use confusion matrix to solve this problem. Confusion matrix is a specific table layout that allows visualization of the performance of an algorithm, each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The Figure 7 shows a classic confusion matrix. In the matrix, we only focus on the true positive and true negative, we assign the each outage types which has the highest true positive to the threads that belongs to many categories, and then we assign each outage types which has the lowest true negative to the threads that don't belong to any category. Then we get a reasonable prediction. Figure 8 shows every outage type's accuracy of true positive and true negative.

4. RESULT

4.1 Outage Types Distribution of Each Year

We calculate the percentage for every outage type in every year. Figure 9 shows us that in 2006, the natural disaster and fiber cut occupy the most outage. In 2007 and 2008, the network maintenance occupy the most



	P′ (Predicted)	N′ (Predicted)
P (Actual)	True Positive	False Negative
N (Actual)	False Positive	True Negative

Figure 7: Classic confusion matrix

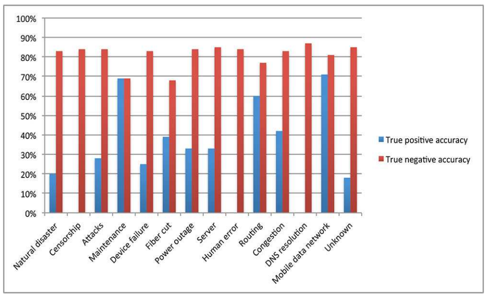


Figure 8: True Positive/Negative accuracy of each outage type

outage. But after 2008, with the development of smart phone, the mobile network outage is increasing. Especially in 2013. the mobile network outage exceeds over 50%. So the trend of outage type in our datasets meets the trend of network development.

4.2 Percentage of Every Outage Type

We calculate the percentage for every outage type from 09/2006-03/2015. Figure 10 shows us that the Mobile Data Network comprises 42%, Maintenance and Fiber cut comprise about 20% and 7%. Therefore these three types of outage domain about 70% in the data set, which means that the majority of outages are users impact. In other words, users play the main role in the major outages. After the issues related to users, the types tends to be more professional which includes Routing, Congestion, and DNS resolution. It shows that those protocols or mechanism in the network are robust relatively. Finally, the natural disaster, censorship, attacks, device failure are less common compared with others. This result tell us that we should pay more attention on the mobile network(wireless) as the mobile end users are rapidly increasing.

4.3 Real-Time Outage Type Predict System

We implement the real-time outage type predict system, which means when one posts a network outage to the mailing list, we can get the email content and analyze it. Then we predict the possible cause of the outage type using our program. Finally, the program posts the analysis result to our website[10] and updates the outage type distribution pie figure in real time. The whole process is automatically. Here is our demo video https://youtu.be/tlc_QVkEqV4

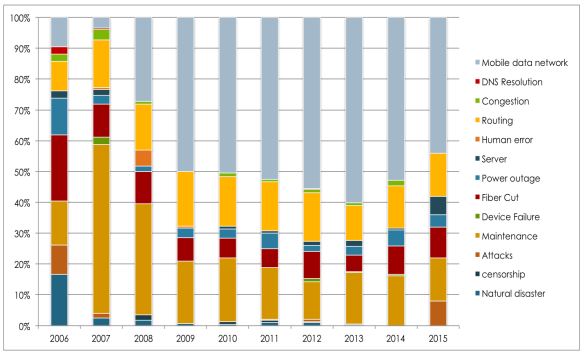


Figure 9: Outage Types Distribution of Each Year

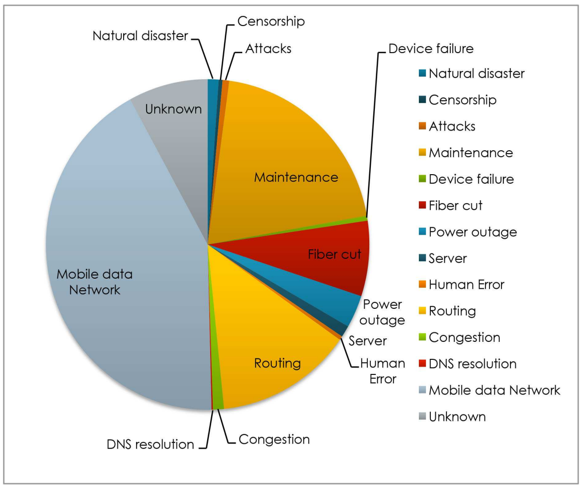


Figure 10: Percentage of Every Outage Type

4.3.1 *How we do it*

We firstly monitor the emails coming from outages.org (subscribers received email from outages.org when a new post comes up). When we get the email subject and content, we import those data to our program, where we have integrated the data preprocessing and classification method. Then the program will return us the predicted result and post associated information to our website.

4.3.2 *What we show on the website*

If the email content includes info about traceroute, then we will extract them and show them separately. And we will also combine all 2015's outages archives to draw a pie figure of the outage type distribution till the present time.

5. CONCLUSION

In this paper, we have integrated the posts with the same subject to the same thread in the Outage mailing list. Furthermore, we have extracted and omitted the unessential data information and maintain the important data information for our classifier by using a stop words list that was obtained from the SMART information retrieval system[5]. We manually increased the amount of words in the stop words list from 571 to 1514, labeled 315 threads as our training data and used python Natural Language Toolkit(NLTK) to lemmatiz the data and improve the classification. Then, we tried out TF-IDF and found out that a word with high td-idf value within a thread doesn't mean the word is useful for the classification; as the result, we imported name-word library and city-word library of python Natural Language Toolkit(NLTK) to avoid this situation. After this, we generated two-dimensional matrix and obtained the useful data for our classification and got the classification result. The last step is to predict a outage type for a new thread based on a reasonable classifier. We use a Fast Linear SVM Solvers for Semi-supervised Learning called svmlin, it is well-suited to classification problems involving a large number of examples and features. It is primarily written for sparse datasets (number of non-zero features in an example is typically small). We use Deterministic Annealing (DA) algorithm for Semi-supervised Linear L2-SVMs. The results of our analysis are substantially in accordance with the threads'(data) actual outage types. These help us to conclude some features of outage causes.

5.1 Feature of Outage Causes

5.1.1 *Mobile network issues are increasing*

With the mobile users increasing these few years, we find the mobile network issues are rising. It's consistent with our result getting by analyzing the mailing list of outages. From 2006 to 2015, the mobile data issues contributes about 40% of whole outages. Those keywords are always related to AT&T, Verizon, Level 3 and some mobile application misconfiguration problems.

5.1.2 *Common outage types are always related with users*

Compared with the outage caused by routing, congestion and DNS resolution, the common outages are always related with users, which shows the outages that are easily to be aware by user are more likely to be reported. Other types, such as intentional types(censorship, attacks) and natural disasters are only a little amount throughout the dataset.

6. FUTURE PLANS

Beside the tasks that we implemented so far, there are more things that we can do such as to analyze keywords with associated outage type in advance and integrate data based on subjects instead of threads to compare the accuracy.

6.1 Analyzing keywords with associated outage type

Since we have all labeled data, we can analyze keywords again to see what outage types that the keywords are frequently distributed to. Then, we can know what most common outage types a certain Internet Service Provider(ISP) or Content provider has and to do the research to understand why and how to improve it. Figure 11, 12 show two examples one for Facebook, another for Sprint.

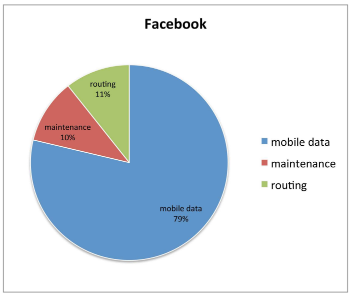


Figure 11: Most common outage types related to Facebook(Content provider)

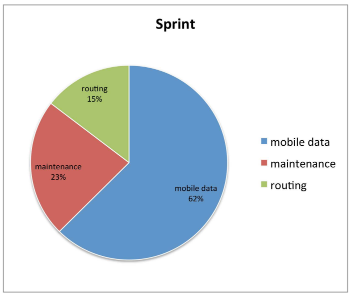


Figure 12: Most common outage types related to Sprint (ISP)

6.2 Integrate data based on subjects instead of threads

Despite the number of threads is very evenly distributed in this ten years. It does not guarantee that using threads as our based unit for analyzing Outages mailing list will always come up with the best accuracy. As a result, we can try out by using subjects as our based unit and compare it with threads to see the results.

7. 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. *Outage (planned & unplanned) reporting.* Retrieved from https://puck.nether.net/mailman/listinfo/outages

[5] J. J. Rocchio. *Relevance feedback in information retrieval*, 1971. Retrieved from http://jmlr.org/papers/volume5/lewis04a/a11-smart-stop-list/english.stop

[6] J. Ramos. *Using TF-IDF to determine word relevance in document queries* In Proc. International Conference on Machine Learning (ICML), 2003.

[7] virendra.rode@outages.org *Internet outages mailing list*, 2006. Retrieved from https://puck.nether.net/mailman/listinfo/outages

[8] J.R.Landis,G.G.Koch,etal. *The measurement of observer agreement for categorical data.* biometrics, 1977.

[9] *Wikipedia - Bag-of-words model* Retrieved from http://en.wikipedia.org/wiki/Bag-of-words_model

[10] *Internet outages analysis* http://zhuguanyu.github.io/fundamental_of_network/realtime/