A Thesis submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Technology

in

Information Technology

by

Vaibhav Vinay Ranshoor

171080067

VEERMATA JIJABAI TECHNOLOGICAL INSTITUTE,

Mumbai University, Mumbai,

H.R. Mahajani Road,

Mumbai - 400019, India.

February 2019

Declaration

I hereby declare that the thesis submitted by me to V.J.T.I College, Mumbai University, Mumbai, 400019 in partial fulfillment of the requirements for the award of **Bachelor of Technology** in **Information Technology** is a bona-fide record of the work carried out by me under the supervision of ***Prof. Pranav Nerurkar***.
I further declare that the work reported in this dissertation, has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma of this institute or of any other institute or University.

Sign: _____

Name & ID. No.: Vaibhav V. Ranshoor, 171080067

Date: _____

VEERMATA JIJABAI TECHNOLOGICAL INSTITUTE

Certificate

This is to certify that the thesis entitled ***Email Analytics*** submitted by ***Vaibhav Vinay Ranshoor*** **(ID. No. 171080067)** to V.J.T.I. , Mumbai University, Mumbai, in partial fulfillment of the requirement for the award of the degree of **Bachelor of Technology** in **Information Technology** is a bona-fide work carried out under my supervision. The dissertation fulfills the requirements as per the regulations of this University and in my opinion meets the necessary standards for submission. The contents of this dissertation have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma and the same is certified.

Supervisor		**Program Chair**	
Signature:	.....	Signature:	.....
Name:	Prof. Pranav Nerurkar	Name:	.....
Date:		Date:	

(Seal of the College)

Abstract

In Email Analytics, our main focus on criminal and civil investigation from large email dataset. It is very difficult to deal with challenging task for investigator due to large size of email dataset. This paper offer an interactive email analytics various to current and manually intensive technique is used for search evidence from large email dataset. In investigation process, many emails are irrelevant to the investigation so it will force investigator to search carefully through email in order to find relevant emails manually. This process is very costly in terms of money and times. To help to investigation process. We combine Elasticsearch, Logstash and Kibana for data storing, data preprocessing, data visualization and data analytics and displaying results. In this process reduce the number of email which are irrelevant for investigation. It shows the relationship between them and also analyzing the email corpus based on topic relation using text mining.

Contents

List of Figures

Chapter 1

Introduction

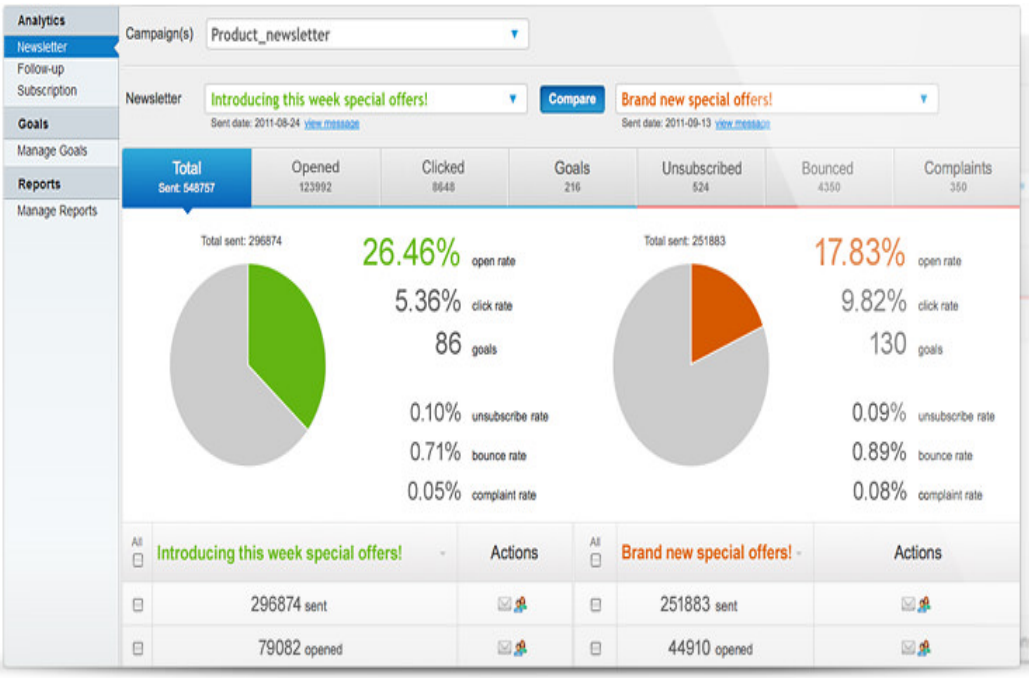
This thesis is written for the project which is under development known as **Email Analytics**. In Email Analytics large email dataset is received and generated. This Email data set is represent technique to discovery of evidence and information in investigation from a large email dataset. So in any large email dataset to prevent the investigator form conducting a manual search. There is a need to reduces effort and saves a lot of business time to automate such activities.

Chapter 2

Email Communication

Email is most of your day to day communication today. The surprisingly fast acceptance of the communication medium. This form of communication has easy to use and costing virtually nothing per message. In the digital age, people use written communication far more than ever before. In fact, email communication is not only used instead of letter writing, it has also replaced telephone calls in many situations and in professional environments. In this book Visualization analysis and Design (Tamara Munzner 2015) Tamara Munzner given explanation about the visual analytics when the exact questions are not known. Email analytics ability to find the human pattern, trends and anomalies.it is very difficult to investigation when content of emails are change.

Email analytics also analyzing the email corpus based on topic relation using text mining. In text summarization a large collections of emails are transformed to a reduced and compact email dataset, which represents the digest of the original email collections. This can be done using topic modeling algorithm. A summarized email helps in understanding the gist of the large email corpus quickly and also save a lot of time by avoiding reading of each individual email in a large email corpus.


Analytics
Newsletter
Follow-up
Subscription
Goals
Manage Goals
Reports
Manage Reports
Campaign(s)
Product_newsletter
Newsletter
Introducing this week special offers!
Sent date: 2011-08-24 view message
Compare
Brand new special offers!
Sent date: 2011-09-13 view message
Total
Sent: 548757
Opened
123992
Clicked
8648
Goals
216
Unsubscribed
524
Bounced
4350
Complaints
350
Total sent: 296874
26.46% open rate
5.36% click rate
86 goals
0.10% unsubscribe rate
0.71% bounce rate
0.05% complaint rate
Total sent: 251883
17.83% open rate
9.82% click rate
130 goals
0.09% unsubscribe rate
0.89% bounce rate
0.08% complaint rate
All
Introducing this week special offers!
Actions
296874 sent
79082 opened
All
Brand new special offers!
Actions
251883 sent
44910 opened

Chapter 3

Objectives

Now a days investigation of email through keyword of both headers and contents of email using methods. Still we are unclear with best keywords for searching the emails in the result. This methodology provides the best keyword for searching the emails in the result. It also reduce the number of emails from large dataset. Data visualization provides relationships in the context of the data, finding human pattern, trends and anomalies easier. Topic modeling provides summarization a large collections of emails are transformed to a reduced and compact email dataset

Chapter 4

Challenges of Email Analytics

First, the data sets are very large and are growing rapidly which provides a challenge to finding relevant information. Second, our interviews revealed a lack of a good set of investigative tools to deal with many of the issues created by large email data sets. These issues include:

- Reducing the size of keywords search results.
- Removing duplicate,irrelevant or unimportant emails from large email datasets.
- Discovering inconsistency in the email data.
- Inability to summarize search results or different subsets of emails data.
- Finding indirect connections between email accounts.

Currently, In the market right now there are no specific techniques or tools to automate this process. The main reason being that the problem which we have to solve is very specific to an organization. email analytics is not providing 100 percent accuracy but it will be enough to automate the process.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

Here I have introduced a new era of analytics that is E-mail Analytics by proposing a new methodology of searching and mining of useful E-mails from large email datasets. Analytics over email dataset makes easier to investigate for investigators to identify hidden relationships and anomalies within the email datasets. This will improve and speed up the results of the investigation process. To find relevant emails, entities and correspondents in the email data sets, investigators found new and interactive technique of visuals which helps them in their decision making. Once these emails are found relevant which were hidden from the initial search through our search result reduction techniques can be brought back into final detailed examination. My study and approach demonstrated visual interaction which can reduce the size of results set, so that the remaining emails can be examined in detail to find relevant emails through investigation.

5.2 Future work

The agenda is analyzing the topic modeling using the elk stack.E Mails and other text forms of communication such as tweets and text messages present a unique challenge due to their nature of the communication.we are not aware of any search system and indexing that have implemented an efficient methods for creating a new index for a subset of document set from index of full document set

Bibliography

[1] Bernard Kerr. Thread arcs: An email thread visualization. IEEE Symposium on Information Visualization, 2003.

[2] C.Ramasubramanian and R.Ramya. Invest: Intelligent visual email search and triage,dfrws usa 2016-proceedings of the 16th annual usa digital forensics research conference, digital investigation. DFRWS USA 2016, 18, 2016.

[3] John Haggerty, Sheryllynne Haggerty, and Mark Taylor. Forensic triage of email network narratives through visualisation. Information Management and Computer Security, 22, 2014.

[4] John Haggerty, Sheryllynne Haggerty, and Mark Taylor. Enron corpus dataset. Information Management and Computer Security, https://www.cs.cmu.edu/ ./enron/.

[5] Haggerty J, Karran AJ, Lamb DJ, and Taylor M. A framework for the forensic investigation of unstructured email relationship data. International Journal Digital Crime Forensics, 2011.

[6] https://lucene.apache.org/.

[7] https://lucene.apache.org/solr.

[8] https://www.elastic.co/.

[9] Enron Dataset,http://www.cs.cmu.edu/ enron/.

[10] Maguire E. ,Munzner T. Visualization analysis and design. AK Peters visualization series. Boca Raton, FL- CRC Press; 2015.