

# A Framework of Email Cleansing and Mining With Case Study on Image Spamming

Pritha Ghosh\*

Received: 20-December-2014; Revised: 18-January-2015; Accepted: 20-January-2015  
©2014 ACCENTS

## Abstract

*“EMAIL CLEANSING” deals with process of eliminating irrelevant non-text data (it includes header, signature, quotation and program code filtering) and transforming relevant text data into canonical form (which includes word, sentence and paragraph normalization). Many text mining applications need to take emails as input. Email data is usually noisy and thus it is necessary to clean it before mining. Email text mining is one of the major parts of email processing. The main purpose of email text mining are Statistical Learning, determining the importance of the email, determine whether the email is spam or not etc. In this paper we are going to address the issue of email cleansing for text filtering as well as spamming based upon text filtering and image filtering.*

## Keywords

*Email Data cleansing; Email Data Mining; Email Processing; Statistical Learning; Image filtering.*

## 1. Introduction

Electronic mail, commonly called email or e-mail, is a method of exchanging digital messages from an author to one or more recipients. With constant increase in the amount of electronic data, the need for tools and techniques to clean and analyse massive data sets has grown rapidly. Email plays an important role in communication. These techniques allow people to communicate worldwide. This vast amount of communication data may potentially contain information, but like any other data sets, it must be cleaned before it is analysed. According to the survey of Radicati group [13] from April 2010, there are

about 1.9 billion users of email worldwide. With the increasing popularity of email, it becomes an important form of communication for many computer users, for both legitimate and illegitimate activities.

The impact of electronic mail in our daily lives is now more obvious than ever before. Each minute, millions and millions of plain-text or enriched messages are being sent and received around the world. Some of which are read with extra care and at the same time, many of them are deleted with obvious disinterest. As the Internet grew, electronic mail has not only turned into a vital tool for our work, but also an important means of interpersonal communication. In professional life, e-mail has played a vital role in Team organization, project management, information exchange (Ducheneaut & Bellotti, 2001) [10], decision making, and client support. E-mail has made personal communication significantly easier as it offered instant messaging with minimum cost. People from all over the world are able to exchange opinions and information with ease that it made e-mail the second most popular channel of communications after voice (Clark, 2003)[7].

Features that made e-mail so popular are the rapidity of communication, the minimum cost, and the fact that it is remarkably easy to use. An advantage over voice communication (e.g. phone) is that it is asynchronous, meaning that there is no need for both sides of communication to be online or in front of a computer at the same time. Unfortunately, e-mail could not escape the curse of information overload. Loads and loads of incoming messages (some extremely important, other simply junk) have made handling of electronic mail, a tedious task.

Today, an average e-mail user receives about 100 or 200 messages per day and in a recent research, IDC<sup>1</sup> predicts that by year 2006, e-mail traffic will be about 60 billion messages per day worldwide (Collins, 2002)[8]. Nowadays, people are struggling to separate important messages that demands

---

\*Author for correspondence

Pritha Ghosh, Department of Computer Science and Engineering JIS Group of Colleges Kolkata.

immediate attention from the mound, and large companies are investing money in order to maintain e-mail centres with personnel dedicated to answer client requests and queries sent by e-mails. Additionally, the problem of spam messaging has grown at a level that it is now considered an industry problem. It costs billions of dollars (Rock Bridge Associates, Inc., 2004) as it takes up bandwidth, clutters in-boxes, and occupies employees who are receiving them. Moreover, the content of many spam messages is unsuitable for children (e.g. pornographic). Hence there is a need to gear up better cleansing, mining techniques.

## **2. Related Works**

### **2.1 Language Processing**

Sentence boundary detection, word normalization, case restoration, spelling error correction, and many other related issues have been investigated intensively in field of natural language processing, but mostly as separated issues [1][3].

### **2.2 Sentence Boundary Detection**

Palmer and Hearst proposed a neural network model to determine whether a period in a sentence is the ending mark of the sentence, an abbreviation, or both [14]. They utilized the part of speech probabilities of tokens surrounding the period as information for disambiguation.

### **2.3 Case Restoration**

Lita et al. proposed employing a language modelling approach to address the issue of case restoration [16]. They defined four classes for word casing namely, all lowercase, first letter uppercase, and all letters uppercase and mixed case, and formalized the problem as that of assigning the class labels to words in natural language texts. They then used the n-gram model to calculate the probability scores of the assignments. Mikheev made use of not only local information but also global information in a document in case restoration [15].

### **2.4 Spelling Error Correction**

Spelling error correction can be formalized as a word sense disambiguation problem. Thus the goal then becomes to select a correct word from a set of confusion words, e.g., {to, too, two} in a particular context. For example, Golding and Roth proposed a statistical learning method to address this issue [12]. The problem can also be formalized as data conversion using the noise channel model from

Information Theory. The source model can be built as n-gram language model and the channel model can be constructed with confusing words measured by edit distance. For example, Mayes et al., Church and Gale [6] [11], Brill and Moore [2] developed techniques for confusing words calculation.

### **2.5 Word Normalization**

Sproat et al. investigated normalization of non-standard words in texts, including numbers, abbreviations, dates, currency amounts, and acronyms [17]. They define taxonomy of non-standard words and apply n-gram language models, decision trees, and weighted finite-state transducers to the normalization.

## **3. Types of Problems Handled**

We have constructed a mailing system by virtue of which we are able to transfer emails among users within a local-hosting networks. And based on the mails exchanged, the mails are cleansed and mined to respective folders.

Following are the cleansing procedures performed on the emails:

- Entire email data is scanned.
- Initially, case restoration is performed upon the email data.
- After performing case restoration problem, sentence boundary detection is performed in which end of sentence is detected and corresponding words first letter is capitalized.
- After performing sentence boundary detection, word normalization is performed in which normalization of non-standard words in texts, including numbers, abbreviations, dates, currency amounts, and acronyms.
- After performing word normalization, paragraph normalization is performed in which normalization of non-standard words in texts, including numbers, abbreviations, dates, currency amounts, and acronyms are filtered [4][5].
- Lastly, image filtering is performed, which is a new concept in the field of email. Image attached in email are contiguous to be spam. Hence proper filtering of image should be done.
- After all the cleansing the emails are mined too respective folder based on the important

data counts (For example: network, TCP, protocols etc correlated words are counted within a mail and based on that it is mined to its respective network folder).

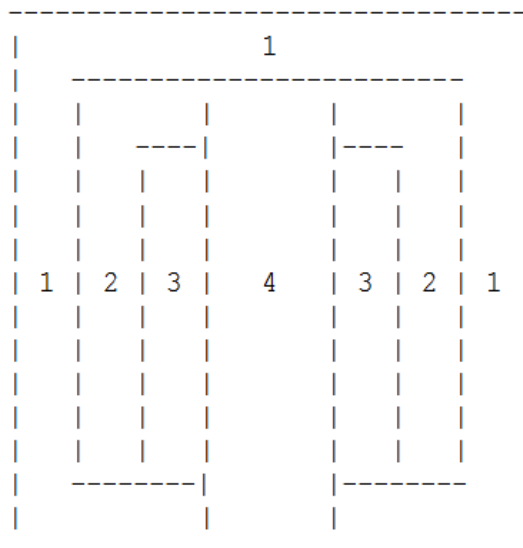
## 4. Image Filtering

Image filtering is a filtering class created to avoid the misuse of image attachment in a mail. Since many image attachments are being utilized for sending pornographic or vulgar images which embarrasses the receiver in many ways and are considered as a spam mail. Hence this image filtering class is used to detect such attachments and perform necessary action. This class is a nudity filter for images. It tests whether an image is nude or not. Currently it works with PNG, GIF and JPG true colors pictures.

### 4.1 How does it work?

The idea of this class is to scan the image pixel by pixel and test if the colour of each pixel is between a human skin colour range, and it returns the percent of found pixels to the number of pixels in the picture [9]. Before scanning we divide the picture into shapes as in fig 1.

As long as you go into the middle of the picture, the colour score will be higher. The idea behind this calculation; that almost all human pictures are centred and focusing on the person who is in the picture.



**Fig 1: Showing the picture divided into shape**

### 4.2 Ideas for better scanning

- After marking a pixel as "skin" we see how many pixels meet together to create a "piece" of skin. But this calculation will take much memory and time.
- Instead of using one big color range, we split it into small ranges, including the African skin colors range. But this will take much memory and time also.
- The class still needs a better color comparison to make sure that colors like the "Yellow" for example (which comes between the min and max color values of the human skin colors) will not be counted. I think this can be done if we compare Red, Green, and Blue values for the pixel color with the Red, Green, and Blue values for the min and max colors.

### 4.3 Limitations of Image filtering class

- No. calculations for African skin colors.
- No. calculations for non-true colors pictures (e.g. 256 colors).
- No. calculations for animated GIF pictures.

## 5. Algorithm For E-mail Mining

The algorithm extracts all the informations from an email sent by a sender to a receiver and performs cleansing and stores the retrieved information.

**INPUT:** Texts of e-mails with all details in a directory

**OUTPUT:** Cleansed e-mail and mined based on the content of the email.

**Step1:** Get the email as a text file.

**Step2:** Extract receiver's email id.

**Step3:** Extract sender's email id.

**Step4:** Perform cleansing operations.

**Step5:** Formatting the email into a standard format.

**Step6:** Link the database with the formatted mail and check with the datasets present in the database.

**Step7:** After checking with the datasets hits are checked.

**Step8:** Checking of attachments is done.

**Step9:** If image attachments are found image filtering is done.

**Step10:** Connect the application program to MySQL database.

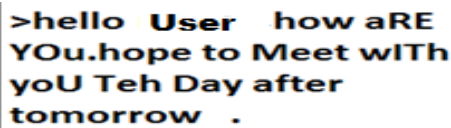
**Step11:** Based on the datasets hit and attachment percent ratio, type of mail is detected and mined to the destined folders of the receiver's mail box.

**Step12:** Update table with the extract details of the mail.

**Step13:** Close the database connection.

## 6. Process and Performance View

On receiving the email from the sender the algorithm extracts the sender's and receiver's email-id and starts cleansing the email which is in textual format into a standard format and links with the database. After linking it checks with the pre-defined datasets present in the database for hits. If no attachments are present then based on the hits the email is mined to the respective folder (network related mails to network folder). If attachment is present and it is and image file it performs image filtering process and mines the mail to respective folder or spam folder. Following are the image of email cleansing:



```
>hello User how aRE  
YOu.hope to Meet wITH  
yoU Teh Day after  
tomorrow .
```

**Fig 2: E-mail message in textual format before cleansing**

**Hello User ,**

**How are you. Hope to meet with you the day after tomorrow.**

**Fig 3: Cleansed E-mail message**

## 7. Conclusion

E-mail is now extremely important for both interpersonal communication as well as professional life. Therefore, there is a need of immediate attention and efficient solutions. Data mining and machine learning has to offer to the clarification of e-mail overload are the intelligent techniques for automation of many e-mail management tasks. E-mail categorization into folders, e-mail answering and summarization and spam filtering are only a few representatives of such tasks that are to be handled carefully. All the applications have been explored repeatedly with very promising results, but spam filtering seems to gather the greater attention of all, probably because of its negative financial impact. It is worth noticing; that many of these applications are extremely demanding in terms of accuracy, mainly because of the information in e-mail data can be significantly important. The algorithm used performs

an efficient job of email cleansing which is utilized for spam detection based on content analysis and is also useful for detection of nudity images or in appropriate images thus helping the cause of detecting spam multimedia or attachment mail messages.

## 8. Future Scope of Work

In this paper, we performed email data cleansing based upon email content analysis as well as image attachments. We light up some future scope of email data cleanings:

- Increasing the efficiency of email data cleaning.
- HTML, which until now was removed in the pre-processing step, could help to give an, at least, semi structured form to the e-mail. Knowledge discovery from structured information is more convenient, and maybe more effort should be made in this direction. For example, we could transform HTML messages into XML using XSL-T patterns.
- Above all efficiency of the best techniques can be taken into account and worked out for better performance.
- Obtain better accuracy level in image filtering in terms of multiple face detection within an image.
- Moreover higher accuracy and lower time complexity of the algorithm can be achieved.

## References

- [1] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. A Maximum Entropy Approach to Natural Language Processing. Computational Linguistics, Vol. 22, 1996, pp39-71.
- [2] E. Brill and R. C. Moore. An Improved Error Model for Noisy Channel Spelling Correction. In Proc. of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'00). 2000, pp286-293.
- [3] Y. Cao, H. Li, and S. Li. Learning and Exploiting Non-Consecutive String Patterns for Information Extraction. Technique Report. 2005.
- [4] V. Carvalho, W. Cohen. Learning to Extract Signature and Reply Lines from Email. In Proc. of Conference on Email and Spam (CEAS'04) 2004.
- [5] H. L. Chieu and H. T. Ng. A Maximum Entropy Approach to Information Extraction from Semi-Structured and Free Text. In Proc. of 18th National Conference on Artificial Intelligence, 2002, pp786-791.

- [6] K. Church and W. Gale. Probability Scoring for Spelling Correction. *Statistics and Computing*, Vol. 1, 1991, pp93-103.
- [7] A. Clark. Pre-processing Very Noisy Text. In *Proc. of Workshop on Shallow Processing of Large Corpora. Corpus Linguistics 2003*, Lancaster. 2003.
- [8] M. Collins. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perception Algorithms. In *Proc. of EMNLP'02*. 2002.
- [9] C. Cortes and V. Vapnik. Support-Vector Networks. *Machine Learning*, Vol. 20, 1995, pp273-297.
- [10] N. Ducheneaut and V. Bellotti. E-mail as Habitat: An Exploration of Embedded Personal Information Management. *Interactions*, Vol. 8, 2001, pp30-38.
- [11] W. A. Gale, K. W. Church, and D. Yarowsky. Discrimination Decisions for 100,000-Dimensional Spaces. *Current Issues in Computational Linguistics: In Honor of Don Walker*. Kluwer Academic Publishers, 1994, pp 429-450.
- [12] A. R. Golding and D. Roth. Applying Winnow to Context-Sensitive Spelling Correction. In *Proc. of the 13th International Conference on Machine Learning (ICML'96)*. 1996, pp182-190.
- [13] <http://www.radicati.com>.
- [14] D. D. Palmer and M. A. Hearst. Adaptive Multilingual Sentence Boundary Disambiguation. *Computational Linguistics*. MIT Press Cambridge, MA, USA. Vol. 23, Issue 2 (June 1997), pages 241–267.
- [15] A. Mikheev. Periods, Capitalized Words, etc. *Computational Linguistics*, 28(3):289-318, 2002.
- [16] L. V. Lita, A. Ittycheriah, S. Roukos, and N. Kambhatla. True casing. In *Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, July 7-12, Sapporo, Japan.
- [17] R. Sproat, A. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards. Normalization of non-standard words. *Computer Speech and Language*, 15 (3) (2001), pp. 287–333.



I was born and raised in Kolkata since the June of 1989 and I am still a citizen of the city till date. Having no other siblings, I am the only child of my parents. I was fortunate to have a convent education from Carmel High School till class 10 before moving onto G.D. Centre of Education for my higher secondary. I pursued my Engineering on Computer Science and technology from a private college at Durgapur from 2007 till 2011 and continued my higher studies i.e. M.Tech in Computer Science and Engineering from Heritage Institute of Technology from 2011 till June 2013. I started working as an Assistant Professor in JIS Group of Colleges and I am still growing with the Institution till today.  
 Email: preet.ghosh89@gmail.com