Machine Learning Email Prediction System (MLEPS)

Taiwo Ayodele, Shikun Zhou, Rinat Khusainov
Department of Electronics and Computer Engineering
University of Portsmouth, United Kingdom

Abstract

Nowadays, email has become one of the most critical personal and business applications and email users would experience serious consequences if email messages could not be available or experience high volume of messages which lead to congestions, overloads and limited storage space coupled with un-structured messages in mail boxes. A few years ago, the means of communication are via letters by post, telegraph, fax, couriers to mention a few but now the focus has changed to a faster means of obtaining quick responses and faster ways of communication-emails. We propose a new framework to help organise and prioritize email better; Machine learning email prediction system (MLEPS). The goal is to organise emails better in mail boxes, prioritise emails based on the focus of the email content. The intelligent email prediction system helps to improve email users' performances, saves time, very effective and efficient tool and is cost effective for businesses and for personal use. The system is evaluated against a corpus of human-judged predictions, reaching satisfactory level of performance.

1. Introduction

We present MLEPS, the machine learning email prediction system. The aim of MLEPS is to help email users save lots of time while checking, reading and searching for email messages reduce email overloads and congestions which are caused as a results of high volume of email messages in mail box, organise mail boxes better and all these makes life easier and improves user's performances and productivity. One of the lines of work developed within MLEPS is the use of machine learning techniques for information management namely, text classification of email messages and determination of messages that needs reply and why.

MLEPS is an automated system that learns to determine whether email messages received in a mail box needs a reply or no action is to be taken. Whittaker and Sidner [1] analyzed the use of email to perform task management, personal archiving, and asynchronous communication and referred to the three as "email overload". They concluded: (1) Users perform a large variety of work-related tasks with email. (2) As a result, users are overwhelmed with the amount of information in their mailbox. A quotation from interviews conducted by Whittaker [1] characterizes some frustrations:
"*Waiting to hear back from another ... employee can mean delays in accomplishing a particular task, which can ... have significant impact on our overall operations. ... it can be critical or just frustrating."*
"One of my pet-peeves is when someone does not get back to me, but I am one of the worst offenders. I get so many emails ... that I cannot keep up."

MLEPS will enable email users to both manage their email inboxes and at the same time manage their time more efficiently. The existing solutions by Tyler et al [2] explained that "regular user of email has, at one time or another, sent a message and wondered, "When will I get a response to this email?" Or, "How long should I wait for a response to this message before taking further action?" This work grew from the belief that an interesting, relatively unexplored aspect of email usage is its implicit timing information". Also Mark et al [3] provided solutions to email reply prediction by assessing date and time in email messages as email containing date and time are time sensitive and may require a reply, and finally used logistic regression with other feature like questions in email message and many more to provide solutions to reply expectation prediction.

2. Previous Work

The email system is one of the most used communication tools in the world. Sproull and Kiesler [4] provide a summary of much of the early work on the social and organizational aspects of email. Here we will focus on work about email reply prediction strategies, as well as research dedicated to alleviating the problem of "email overload and prioritization." Mackay [5] observed that people used email in highly diverse ways, and Whittaker and Sidner [1] extended

this work. They found that in addition to basic communication, email was "overloaded" in the sense of being used for a wide variety of tasks-communication, reminders, contact management, task management, and information storage.

Mackay [5] also noted that people fell into one of two categories in handling their email: *prioritizers* or *achievers*. Prioritizers managed messages as they came in, keeping tight control of their inbox, whereas achievers archived information for later use, making sure they did not miss important messages.
Tyler et al [2] in a recent interview study identified several factors that may influence likelihood of response. These empirical studies were qualitative, generally based on 10 to 30 interviews.

3. Email Reply Management System (ERMS)

Email reply management system is designed to handle incoming email messages for the MLEPS and transferring the mails to the email predictor which then analyse each email fields and contents to determine if they need reply and classify them into various state (*need reply-0, do not need reply-1 and others-2)* based on the sensitivity of the content of the email content.

This is a decision making system that could determine if emails received require a reply and the model is shown in Figure 1. For any given email datasets, there are multiple email conversations and to capture these different conversations, the system assumes that if one email was a reply to the sender's original message, then such a mail may require attention as this may have element of *request*. We used machine learning techniques for finding *interrogative words*, *questions marks*, *most frequent words*, *most used phrases* and embed WorldNet [6] in order build a model that provide a focus to each mail and determine whether email message require a reply.

We implemented a machine learning approach to solve the problem of MLEPS system. Machine learning is learning the theory automatically from the data, model fitting, or learning from examples. It is also an automated extraction of useful information from a body of data by building a good probabilistic model.

3.1. Importance of Machine Learning

Our work involves machine learning because it is the underlying method that enables us to generate high statistical output. These are the importance of machine learning as applied in our work:

- New knowledge about tasks is constantly being discovered by humans. Like vocabulary changes, and there is constant stream of new events in the world. Continuing redesign of a system to conform to new knowledge is impractical, but machine learning methods might be able to tract much of it.

- Environments change over time, and new knowledge is constantly being discovered. A continuous redesign of the systems "by hand" may be difficult. So, machine that can adapt to changing environment would reduce the need for constant redesign.

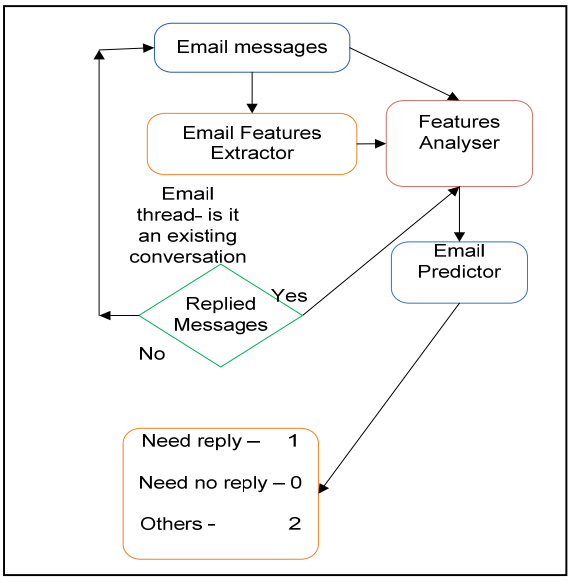

Figure 1. Email Management System

The Figure 1 shows the schematic model of our email management. A schematic model of the architecture for email words and phrases extractions from incoming email messages. Email messages are passed as input data to the feature extractor where all the rare words and frequent words and phrases are extracted and passed to the analyser. The analyser, then check the meaning of each words with the embedded WorldNet [6] for the meaning and our predictor then select most meaningful words that occur often in the mail and build a storage of words and phrases model and store words and phrases learned and their meaning. Our predictor will intelligently choose the most frequent words and meanings to determine the categories that the email messages belong to: *need reply-0, do not need reply-1, others-2.* The predictor's decision is based on human training data that has been learned and begin to be more self intelligent as new mails arrives.

3.2. Machine Learning Email Prediction System (MLEPS)

This is an automated machine learning system that determines if emails received require a reply. We also implemented WorldNet. WorldNet [6] is a large lexical database of English, developed under the direction of George A. Miller. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations.

We embed WorldNet into our email feature analyser as shown in Figure 1 in order to analyse meaning of words and short phrases in email messages and as our analyser finishes with the meaning of words, then the processed email messages are then passed to our propose predictor which then determine the category of the email messages based on the intelligent feedback retrieved from the feature analyser to determine if such a mail belongs to:

- Need reply – 1
- Do not need reply – 0
- Others- 2

The MLEPS learns by example as human participants analysed over 14000 email messages to determine these three aforementioned categories. Further analysis on the methodology used is explicated in section 4. Human participants involved in this research training and testing of this IEPS are from various backgrounds and professions. We have 200 participants from Europe, Asia, Africa, America and Australia and their profession ranges from postgraduate students (Researchers, IT discipline, Banking, medical, Environments, Academia etc), undergraduate students (Art and Science majors), bankers, doctors, business owners, tourism, Air line agents, and many more.

4. MLEPS Participants' Methodology

Two hundred email users at 80 different departments of an IT company and 500 University students were interviewed and observed during their morning reading of emails. Each of them checked the followings:

- Sender's email address
- Subject of each email messages
- Cc/Bcc field
- Content of the emails
- Previous conversation

The email users analysed all their emails ranging from private emails, business emails, to public emails. Each of their email messages were analysed to determine email messages that require reply. The human participant employ heuristic approach to solving the difficulties in prioritising email messages with the following assumptions:

A. Sender's email address: If an email message is from certain people: CEO, Manager, Head of department, debt collector, hospital etc then assume it may require a reply and assign a score of 1 to this on the scoring board provide with the predictor system

B. Subject of each email messages: If the subject of the email messages are similar or almost related to the phrases stored in the database of words and phrases on our system- *please reply soon, let me hear from you, Is there any news today?, Are we on the same project or not.* A score of 3 is allocated to the scoring board.

C. Cc/Bcc Filed: If a mail is Cc or Bcc to others, such a mail may require attention because mail copied to others may be as a result of a group project or task that an individual or groups need to be aware of or act upon as soon as possible. Such a mail may require reply. A score of 2 is allocated to the scoring board.

D. Email content: If a mail contain words and phrases such as: *interrogative words- could, when, where, how? Rarely used phrases: meeting at noon, is it alright, because of the yearly budget, based on what we saw, etc.* such mail may denote a request and may denote a reply. A score of 3 is allocated to this area because content of email messages focus on what the email is all about and that is why human participant spent more time in analysing this area of research.

E. Previous conversation: An existing email conversation within the same subject indicated that majority of such messages require attention. A score of 1 is assigned to the scoring board making a total of 10 points.

Based on the results of these interviews, and testing, our MLEPS predictor is trained to learn the features mentioned above and be to know the type of categories each email received belongs to and be more intelligent as times goes on when received new messages.

5. Machine Learning Techniques

Email messages in mail boxes could become large amount of data that could have hidden correlations and may be hard to find specific emails messages in mail box after a long time. We experiment with machine learning techniques to learn and be self knowledgeable about email features namely:

- *sender's email address (domain from where this email is coming from)*
- *previous email conversation-which may suggest any request made previously*
- *subject field for any phrases that suggest interrogation or statements of commitment*
- *attachment found in email messages.etc*

Our technique is capable of learning email features that could be used to determine whether an email require a reply and is capable of becoming more intelligent when it receives a new email with different format ranging from public email, e-commerce, private and business emails. The technique keeps learning and that makes it more efficient and effective learning approach without any supervision.

6. Evaluations and Results

The following observations were made from the initial interviews. The Table 1 shows detail results:

- All users scanned new messages several times in order to read and determine the most important messages and at the same time categorize them.
- Messages related to events, meetings, venues are considered as important.
- Replies are important as they often contain a solution to a problem posted by the recipient and there are usually elements of request.
- Subject field of email messages do give a clue about what the mail is all about.
- For most users, carbon copies and Blind copies were judged by their receivers as important as other messages.
- Messages containing interrogative words are many and most recipients consider such messages as need reply.

Details about individual users' number of messages are displayed in Table 1.

Email messages are categorised by our intelligent email reply prediction system to require a reply or not in three mail categories as shown in Figure 1. The three possible predictions for the email messages are:

- **Need reply:** Email messages that are categorised to need a reply indicated that such a mail passed the threshold set by human analyser. The threshold value assigned for email messages that require a reply is 7 and any messages that score 7 or above out of total score of 10 will be assigned the tag "*need reply -1"*.
- **Do not need reply:** Messages that scoreless that the threshold value 7 will be assigned a tag "*Do not need reply- 0"*.
- **Others:** Messages that could not belong to either of the categories above will be categorised here. These email messages that are in this categories are: *email messages from friends that does not require any urgency, auto respond reply messages, advertisement email messages, junk emails, email messages with email address: nonreply@myname.com, informative emails etc.*

Intelligent email reply prediction system was evaluated using precision and recall over 14,000 email Enron datasets from over 200 email boxes owned by 350 people from Enron Corpus [7] as shown below:

$$\text{Recall} = \frac{\text{group found and correct (needs reply)}}{\text{total group correct (rightly predicted)}}$$

$$\text{Precision} = \frac{\text{group found and correct (needs reply)}}{\text{total group found (Total email found)}}$$

The Table 1 shows the results of interviews perform on a small scale basis for the first and second interviews. As seen on the table below, MLEPS has reduced email users' time spent on un-necessary browsing through thousands of email messages as this solution has made the mail box well organised and well structured.

Table 1. Participant Interview Output

Participants	Emails per day	Emails that req. reply	Email that does not req. reply	Others	% of Time saved
Uni. Staff	20-60	15	30	15	25%
Researchers	15	4	8	3	27%
Medical Staff	30-65	20	26	9	31%
IT and Eng Researchers	80-120	35	60	25	29%
Financial Researchers	200-270	105	35	30	39%
Students	22-40	12	25	3	30%
Business Workers	100-296	124	160	12	42%
Private and Public agents	102-208	180	20	8	87%

Email messages are well prioritized and therefore overload is reduced and high volume of email messages are well controlled by the proposed MLEPS. The system performs better than the existing email prediction systems. The Figure 2 shows the prototype intelligent email prediction system.

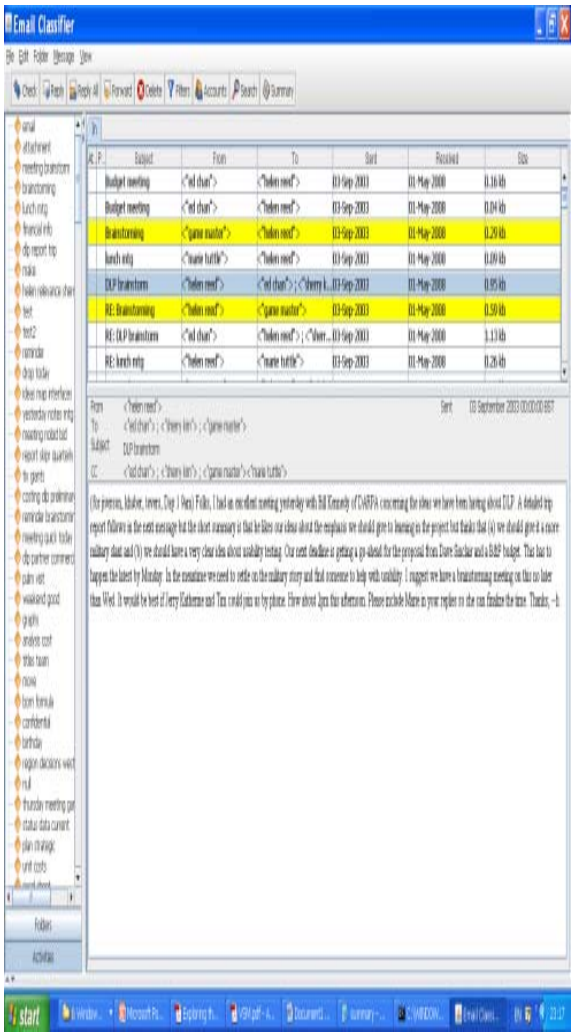

Figure 2. IEPS Prototype

The Figure 2 shows our intelligent email prediction system and email messages tagged as yellow are the email messages that need reply.

6. Conclusion and Future Work

In this paper, we study how to generate accurate measures to determine the mail that require a reply and the one that does not require reply. We analyse the features of emails and study email conversation structure, MLEPS system learns from human participant categorised data tests, which we maintain that this area of research has not been sufficiently investigated in previous research on intelligent email prediction system. We build a novel structure: email management prediction system (EMPS), interrogative words, mails from specific domains and many more. Our future plan includes improving the EMPS with more sophisticated linguistic analysis. In order to verify the generality of our findings, we are working on evaluating our methods with different real life datasets: creating the gold standard for a large real-datasets.

7. References

[1] Whittaker, S., Sidner, C. *Email overload: exploring personal information anagement of email*. in *In Proceedings of CHI'96 Conference on Computer Human Interaction.* 1996. New York: ACM Press.

[2] Tyler, J.R., Tang, J. C. *When can i expect an email response? a study of rhythms in email usage*. in *In Proceedings of the Eighth Conference on European Conference on Computer Supported Cooperative Work.* 2003: Kluwer Academic Publishers.

[3] Mark, D., John, B., Fernando, P. *Reply Expectation Prediction for Email Management.* in *In the 2nd Conference on Email and Anti-Spam (CEAS 2005).* 2005. Stanford University, California, USA.

[4] Sproull, L., Kiesler, S. B., *Connections: New ways of working in the networked organization.* 1991, Cambridge, MA, USA: MIT Press. 212.

[5] Mackay, W.E., *Diversity in the use of electronic mail: a preliminary inquiry*, in *in ACM Transactions on Information Systems (TOIS).* 1998, ACM Press: New York, NY, USA. p. 380-397.

[6] Miller, G.A., Fellbaum, C., Tengi, R., Wakefield, P., Langone, H. *WordNet.* 2009 09/01/2009 [cited 2008 12/06/2008]; Available from: http://wordnet.princeton.edu/.

[7] Bryan, K.Y. *The Enron corpus: A new dataset for email classification research*. 2004.