Email Grouping and Summarization: An Unsupervised Learning Technique

Taiwo Ayodele Shikun Zhou Rinat Khusainov
Department of Electronics and Computer Engineering
University of Portsmouth, United Kingdom
{taiwo.ayodele, shikun.zhou, rinat.khusainov} @port.ac.uk

Abstract

This paper presents the design and implementation of a system to group and summarize email messages. The system exploits the subject and content of email messages to classify emails based on users' activities and auto generate summaries of each incoming messages. Our framework solves the problem of email overload, congestion, difficulties in prioritizing and successfully processing of contents of new incoming messages and difficulties in finding previously archived messages in the mail box by providing a system that groups emails based on users' activities, and providing summaries of emails.

1. Introduction

One of the annoying things is when you are away from the office or home for a while and do not have access to your emails while on holiday or on a trip for a week or more and come back to realize that you have hundreds of email messages waiting for you. The question of where to start from or how to decide which email needs attention comes in. Where will you start from and which email needs respond and how will you maximize your time now that you are back, and have lots of work load in your desk to deal with. This is where our solution to reducing email overload, email congestions and high volume of email originated from. Electronic Emails are parts of everyday life. Personal computer users use emails to communicate with friends, families, e-businesses and colleagues allowing ease of communications. Even with effective methods of controlling spam, the tide of potential irrelevant messages continues to rise. Emails to some people serves as archival tools, many users never discard messages because their information contents might be useful at a later date - for example, as a reminder of upcoming events and outstanding issues. Also, a paper by Schuff et al [1] states that "Emails are widely used to synchronize real-time communication, which is inconsistent with its primary goals". Email messages are designed to be sent, accumulate in repository and be periodically collected and read by receipt, which lends itself to the details of a vacation or a meeting's upcoming agenda.

Since most people rely on emails for efficiency and effectiveness of communication, mail boxes may become congested. Messages range from static organization knowledge to conversations with such a broad horizon of messages. Users may find it difficult to prioritize and successfully process the contents of new incoming messages. Also it may be difficult to find a previously archived message in the mail box.

Kushmerick [2] stated that "the ubiquity of email and its convenience as knowledge management tools make it unlikely that users' behavior will change as falling bandwidth and disk storage prices further reduce the incentive to steer away from using email as a document storage system". At this stage, new effective method for managing information in email, reducing email overloads is developed by grouping emails based on users' activities, and providing summarization of emails in this research.

2. Related Work

There is little exploration into the problems of categorizing and grouping emails into folders but less work in classification of emails based on the activities of users -based on what the users do.

One of the common existing methods used for email classifications is to archived messages into folders with a view of reducing the number of information objects a user must process at any given time. This is a manual classification solution, however, this is an insufficient solution as folder names are not necessarily a true reflections of their content and their creation and maintenance can impose a significant burden on the user [1]. There are some examples of existing email classifiers and some of them are:

978-0-7695-3507-4/08 $25.00 © 2008 IEEE
DOI 10.1109/CSIE.2009.298

- Ishmail [3]: It automatically sort email messages into folders and order them by importance.

- Commercial email clients [4]: Most popular commercial email clients like Procmail, Eudora, Mozilla Thunderbird, Microsoft Outlook and Outlook Express also supports message filing according to user-defined rule sets.
- IBM's MailCat [5]: It adapts dynamically to a user's observed mail-filling habits and provides a list of three folders most likely to be appropriate for a given message.
- Magi [6]: This records each email interaction and uses a learning algorithm to classify new messages based on the user's prior behavior

Moreover, a rule-based system as explained by Schuff et al [1] can provide straight forward way to semi automate email classification and such system require the users to define a set of instructions for the email application to sort incoming messages into folders and order them by importance. The disadvantages of rule-based system are that they are challenging for non technical users because writing the rules require some level of programming experience. Bifrost an email classifier and a prototype email management system [3] avoids this difficulty by letting user define all filtering rules with a simple graphical interface.

Terry et al [4] proposed a new approach by automatically assessing incoming messages and making recommendations before emails reach the user's inbox, so the priority system classifies each messages as of either high or low importance based on its expected utility to the user. While Kushmerick [2] designed a system that automatically identifies messages belonging to the same structure activity, an electronic commerce transaction, thereby providing a high-level view that supports the use of email as a task manager and in summary Boone [6] describe Re: Agent system group similar messages based on existing folder structure provided by the user while it learns concept and decision policies for future message classification based on these folders examples.

Also, the implementation of existing summarizers and the techniques that are used in most summarization systems are the use of false positive regular expressions and also relying on existing software to find names of people and companies mentioned in certain messages. Also some implemented the use of gold standards as references. Zhou et al [7] states that human-written summaries usually make up the gold standards. And due to the complex structure of the email dialogue, the summary itself exhibits some discourse structure, necessitating such reader guidance phrases such as "for the … question," "on the … subject," "regarding …," "later in the same email," etc., to direct and refocus the reader's attention.

2.1 Survey

AOL research source investigated email as the most frequent used communication tool as shown in Figure 1 below. As email services advance, increasing volumes of email can flood users' mail boxes and can lead to congestion problem. Users will not be able to view contents of incoming mails and may find it difficult to find important mails in their mail box. Figure 1 shows more survey results.

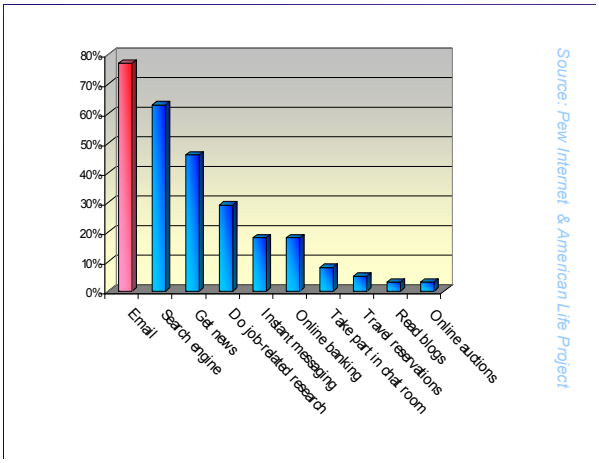


Figure 1: Emails as Most used Communication tools

Our research survey in combination with other survey achieved that most people use email as means of communication nowadays in comparison to other means of communications; Letters, fax, telegraph and courier service. Figure 1 shows that more than 80% use email as the major means of communication: flight booking and confirmations, online purchases, online invoices sent via emails, communication within organizations, businesses and many more are majorly via emails.

3. Contributions

Our contribution in this area of research is "consideration of highest frequencies of words in email messages, with selection of the sentences that contains the most frequent words and re-arranging the sentences in orders that generate a good summary and concisely group the email into activities based on the content of the message" using build in users' favorite dictionary

as well as vocabulary model- *the algorithm will use this to check up the meaning of words*. So comparing to other techniques in [1, 8, 9], this is more suitable for real time email client system because of the efficiency and good performances.

4. Methodology

Our email grouping system (EGS) as well as the email summarization system (ESS) uses heuristic approach with unsupervised machine learning techniques. We extract phrases "*we have a **meeting in the conference room tomorrow by 2pm**. Hope all of the networking them will be there"* and vocabularies from email messages, subject of email messages and check up the vocabularies with users' favorite dictionary that was developed. We build an activity model through which emails that pass through the email grouping system (EGS) are analyzed- *check the subject as well as the message contents for phrases and vocabularies used* and determine what the email is about after checking the vocabulary model with the dictionary and determine the correct *activity* based on its intelligent knowledge. When such activities are generated, then our EGS system will determine whether such email belong to an existing group or a new group has to be created. If it belongs to an existing group, then this becomes a sub-activity thread in this group but if it is a new activity, then a new group will be created.

We also implemented unsupervised machine learning approach with our email *summarization system (*ESS). We exploit most frequent words and phrases in email messages with combination of sentences that contains these words or phrases. We also train our email summarization algorithm on various types of email messages as we receive more emails that ranges from personal emails, public, business, e-commerce emails etc, the algorithm becomes more intelligent. If an email is about flight booking confirmation, our ESS will extract subject field and check for e-ticket as well as the content of the message- Airline reference number. If these are found, then our algorithm knows that this is about flight booking, so, it goes on further to extract the flight number, departure and arrival time to make its summary. Our summarization system is able to learn and continue to learn when new email of any content types are received as our algorithm keeps increasing its knowledge and becomes more intelligent as new ideas and new vocabularies are found.

5. Evaluation and Results

Our email grouping and summarization are evaluated using precision and recall. Enron Corpus was used and 4000 messages were downloaded from 100 mailboxes owned by 90 people. Our algorithm calculates precision and recall as:

$$\text{Recall} = \frac{\text{group found and correct}}{\text{group correct}}$$

$$\text{Precision} = \frac{\text{group found and correct}}{\text{total group found}}$$

Our unsupervised machine learning approach achieved 98% accuracy in comparison to the gold standard. The evaluation results are explained below.

5.1 Email Grouping

This is an automated classification system where a set of rules are learnt by the proposed classifier and is trained to apply the learnt rules on the incoming email messages as it gets to the mail inbox. This enables the emails to be categorized base on the users' activities. However, the effectiveness and the accuracy of the classifier depend on the correctness of the rules implementation learnt and trained to execute.

The classifier provide a heuristics-based approach to extract common words (repeated words) in the subject of the email as well as the content of the mail and get all words frequencies in the subject, this means the number of times that each word occurred in the content and here we use the stop words to prevent the algorithm to count unnecessary words like "the, a, in, at". The algorithm is explained in figure 2 below:

```
Let the message downloaded be M
If(M downloaded successfully)
 then
{
    1. let FW = most frequent words in the message
    2. iterate over all activities and for each activity say AC

        a. let AFW = common words in activity AC
        b. if(FW is most likely AFC) then
        c. mark the activity AC as the one for
        d. else continue to the next activity
    3. if (AC is not null) then
            a. update the message activity as AC
            b. else create new activity called AC and set
               its AFW to be FW
            c. let AN be the common words of the
               message after it is ordered according to its
               position in the message content or subject
            d. set the name of AC as AN
    4. create a rule that states // machine learning

            a. for each message received that has some
               words like FW
            b. AC is the activity for this message
        }
```

Figure 2: Email Classification Algorithm

The evaluation result of precision and recall result as compared with gold standard are shown in table 1 below.

Table 1: Precision and Recall Result

Correct predicted group	Total Predicted Group Found	Total Emails	Precision	Recall
310	316	320	98.1%	96.9%

This is a classifier sample interface result as shown in Figure 3 after testing with different email datasets including Enron email datasets, and privates email messages.

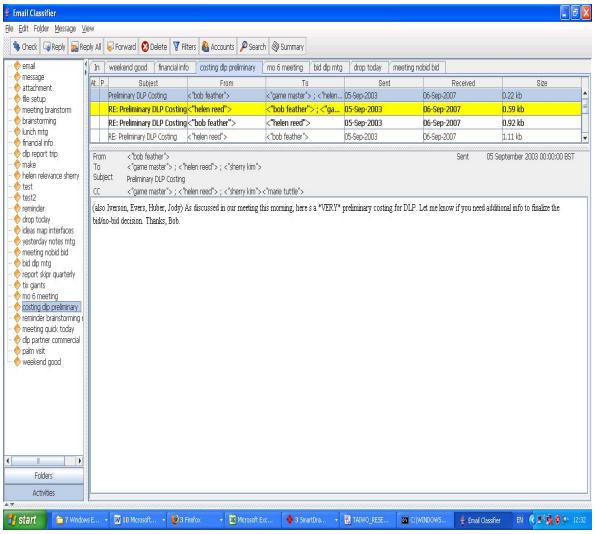


Figure 3: Users' activities are on the left

Here, when the emails are received by the email client, the incoming emails are passed unto the classifier and this then group the emails into activities that users perform. In Figure 3, the incoming email messages are grouped into users' activities and these activities are shown at the left hand side of Figure 3. So, if the same email belongs to the same activities, they tend to form a structured thread and be grouped into the activities based on the content of the email message.

5.2 Email Summarization

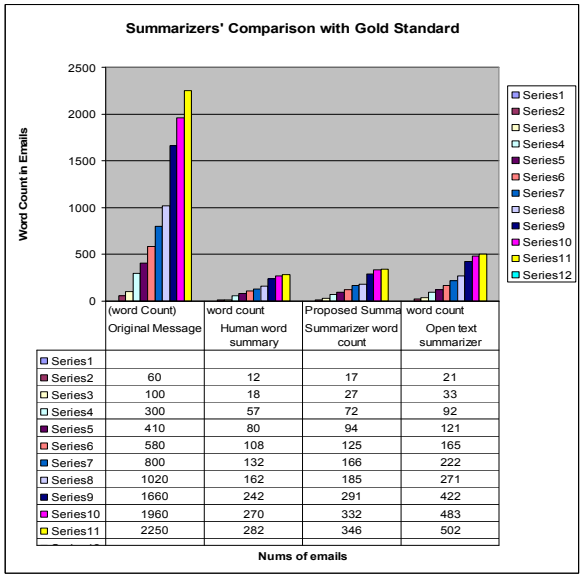
This algorithm extracts important words in email messages so that the summarizer can generate a more useful summary from the message. The algorithm works logically based on the techniques as shown in the Figure 4 below:

Summarization Algorithm

Input: N, M, Msg Output: Sentence list
1). Identify N most frequent words in incoming email messages
2). Select M sentences from email containing most frequent words
3). Order the selected sentences according to their occurrence in the message
4). Output the ordered sentences as summary

Figure 4: Iterative algorithm for summarization

To measure the quality and goodness of the email summaries, gold standards are used as references. Zhou [7] stated that human written summaries make up the gold standards. We evaluate our proposed email summarizer against summaries from human participants and well as open text summary software as shown in Figure 5. While Figure 6 show the original message received and 7 shows the real time summary output with a new design mail client.



	(word Count) Original Message	word count Human word summary	Proposed Summa Summarizer word count	word count Open text summarizer
Series1				
Series2	60	12	17	21
Series3	100	18	27	33
Series4	300	57	72	92
Series5	410	80	94	121
Series6	580	108	125	165
Series7	800	132	166	222
Series8	1020	162	185	271
Series9	1660	242	291	422
Series10	1960	270	332	483
Series11	2250	282	346	502

Figure 5: Email Summary Comparison

We evaluate the total numbers of word in each email and experiment with our proposed summarizer to test how good the summary could be and in the above it is noted that the numbers of words in the original message has been reduced as shown by the various

summarizers and our proposed summary seems to summarize email message better.

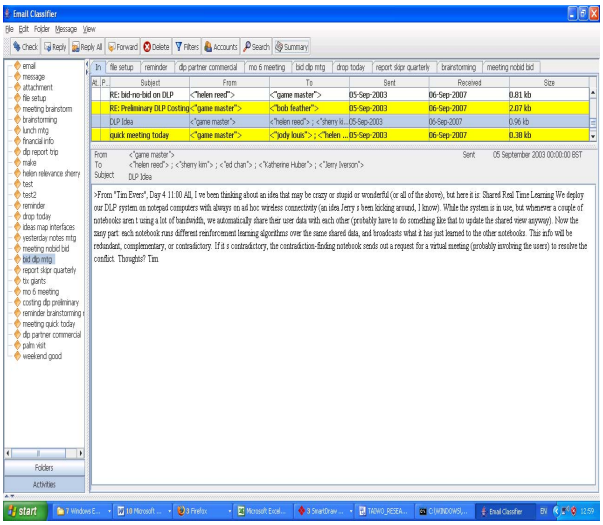


Figure 6: Original Message received

Figure 6 shows an email message received. So, as the mail comes into the mail box, the mails are passed unto the summarizer and the summarizer makes a summary of each mail. The left sides of figure 6 shows the activities of the user and the summary of what the mail is about and if one clicks the activities, the summary of the mail will be shown. Figure 7 below shows the summary of the mail above.

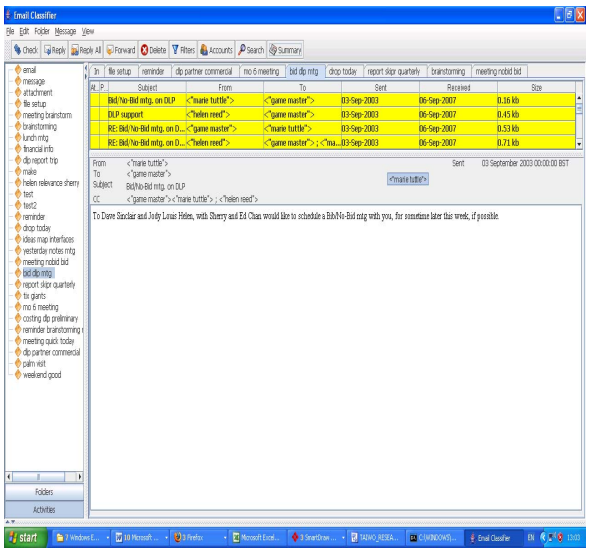


Figure 7: Summary of Original message

The email above has been summarized as shown in Figure 7 by the propose summarizer and gives the concise summary that is meaningful.

6. Conclusions

We have presented an overview of the proposed solutions to extract important words in email messages to provide a better summary than simply running the unprocessed message. As this is another better way of generating useful summaries thus far. Our system also would be able to group emails messages into user's activities and provide a mechanism for emails that needs attention.

We analyze the features of emails and study email conversation structure, users' favorite dictionary, vocabulary model, which we argue and have not been sufficiently investigated in previous research on email classification and summarization. The email grouping system as well as email summarization system relies on a simple algorithm but it is very complex to implement. Yet it appears to work better than other existing approaches.

References

[1]. D. Schuff, O. Turetke, D. Croson, F 2007, 'Managing Email Overload: Solutions and Future Challenges', *IEEE Computer Society, vol. 40, No. 2, pp*. 31-36.

[2]. N. Kushmerick, T. Lau, 2005, '*Automated Email Activity Management: An Unsupervised learning Approach', Proceedings of 10th International Conference on Intelligent User Interfaces*, ACM Press, pp. 67-74.

[3]. J. Helfman, C. Isbell, 1995, 'Ishmail: Immediate Identification of Important Information', AT&T Labs.

[4]. G. Boone, 1998, 'Concept Features in Re: Agent, An Intelligent Email Agent', *Proceedings of 2nd International Conference on autonomous agents,* ACM Press, pp.141-148.

[5]. R.B. Segal, J.O. Kephart, 2002, 'MailCat: An Intelligent Assistant for Organizing Email', *Proceedings of 3rd Annual Conference on Autonomous Agent,* ACM Press, pp. 276-282.

[6]. T. Payne, P. Edwards, 1997, 'Interface Agents that learn: An Investigation of Learning Issues in a Mail Interface', Applied Artificial Intelligence, vol. 11, no. 1, pp1-32.

[7]. L. Zhou, E Hovy, 2005, "On the Summarization of Dynamically Introduced Information: Online Discussions and Blogs", In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, Stanford, CA.

[8]. D. Lam, S. Rohall, C. Schmandt, M. Stern, F 2002, 'Exploiting Email Structure to improve Summarization', A Collaborative User Experience Technical Report (TR2002-02), IBM Watson Research Center.

[9]. S. Whittaker, C. Sider, 1996, 'Email overload: exploring personal information management of email', *CHI '96'*, pp.276-283. ACM Press.