Journal of Engineering and Technology Research Vol.1 (7), pp. 143-151, October, 2009
Available online at http:// www.academicjournals.org/JETR
© 2009 Academic Journals

Full Length Research Paper

Applying machine learning techniques for e-mail management: solution with intelligent e-mail reply prediction

Taiwo Ayodele* and Shikun Zhou

Department of Electronics and Computer Engineering, University of Portsmouth, United Kingdom.

Accepted 17 August, 2009

In today's world, much of our communication is done via e-mail. Many companies and internet users now view e-mail as one of their most critical personal and business applications and would experience serious consequences if their e-mail messages could not be available or experience high volume of messages which lead to congestions, overloads and limited storage space coupled with un-organized e-mail messages. A few years ago, the means of communication are via letter 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, e-mails. We propose a new framework to help organised and prioritized e-mail better; e-mail reply prediction. The goal is to provide concise, highly structured and prioritized e-mails, thus saving the user from browsing through each email one by one and help to save time.

Key words: E-mail reply prediction, e-mail messages, interrogative words, requires reply, questions.

INTRODUCTION

One of the annoying things is when someone does not get back after sending so many e-mail messages to them or when one is waiting to hear back from a friend or colleague at work about completing a particular project which can have a severe impact on the overall operation. This can be frustrating.

E-mail prediction is a method of anticipating if e-mail messages received require a reply or did not require any urgent attention. Our e-mail prediction system will enable e-mail users to both manage their email inboxes and at the same time manage their time more efficiently. Bradley et al. (1996) analyzed that Remembrance Agent (RA) is a program which augments human memory by displaying a list of documents which might be relevant to the user's current context. Unlike most information retrieval systems, the RA runs continuously without user intervention. Its unobtrusive interface allows a user to pursue or ignore the RA's suggestions as desired. This idea was implemented in information retrieval and his approach relies on continuous searches for information that might be of use in its user's current situation. For example, while an engineer reads email about a project the remembrance agent reminds her of project schedules, status reports, and other resources related to the project in question. When she stops reading e-mail and starts editing a file, the RA automatically changes it recommendations accordingly.

The existing solutions by Joshua et al. (2003) 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 e-mail?" 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 e-mail usage is its implicit timing information". Also Mark et al.(2005) provided solutions to e-mail 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 email reply predictions. Other studies have focused on how people save their e-mail, what purposes

*Corresponding author. E-mail: Taiwo.Ayodele@port.ac.uk.

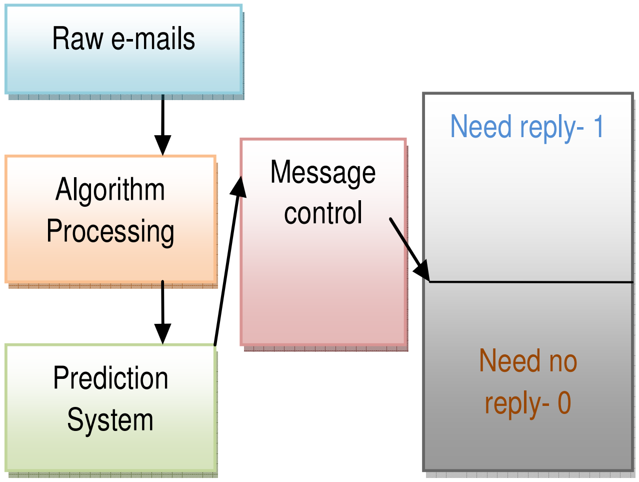


Figure 1. Architecture for words extraction from incoming e-mail.

it serves for them, and its importance as a tool for coordination in everyday life (Laura et al., 2003; Ducheneaut and Belloti, 2001; Mackay, 1998; Sproull and Kiesler, 1991; Mary, 1985; Kraut et al., 1997).

This paper proposes to solve the problem of email prioritization and overload by determining if email received needs reply. Our prediction system provides a better and efficient way of prioritizing email messages as well as provides a new method to email reply prediction.

PREVIOUS WORK

Because email is one of the most used communication tools in the world. Sproull and Kiesler (1991) 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 “e-mail overload and prioritization.” Mackay (1998) observed that people used e-mail in highly diverse ways, and Whittaker and Sidner (1996) extended this work. They found that in addition to basic communication, e-mail was “overloaded” in the sense of being used for a wide variety of tasks-communication, reminders, contact management, task management, and information storage.

Mackay (1998) also noted that people fell into one of two categories in handling their e-mail: 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 and Tang in a recent interview study identified several factors that may influence likelihood of response (Robert et al., 1997). These empirical studies were qualitative, generally based on 10 - 30 interviews.

SYSTEM DESIGN

We used machine learning techniques for finding interrogative words, questions marks, most frequent words, most used phrases with self built dictionary that can determine whether email message require a reply.

We implemented an unsupervised learning approach to solve the problem of email reply predictions. 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.

Importance of unsupervised 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.
- Some tasks cannot be defined well, except by examples and large amounts of data may have hidden relationships and correlations. Only automated approaches may be able to detect these.

Figure 1 shows the schematic diagram of the architecture for e-mail words extraction from incoming email messages for efficient reply prediction proposes in this work.

Our proposed prediction system accept email messages as input data and emails are passed unto our machine learning prediction algorithm system, e-mail header features are obtained from each e-mails and the predictor determines in numeric values the mails that require replies and the emails that does not require replies as shown in Figure 2 below.

While YES is assigned a value 1 which mean it require a reply and NO is assigned a value 0 which means such a mail with this tag does not require a reply.

E-mail reply prediction (ERP)

This is a decision making system that could determine if e-mails received require a reply. For any given e-mail

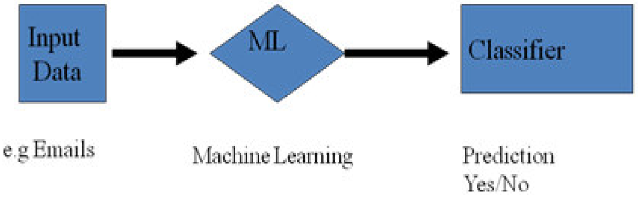


Figure 2. E-mail predictor system.

datasets, there are multiple e-mail conversations and to capture these different conversations, we assume that if one e-mail was a reply to the sender's original message, then such a mail may require attention as this may have element of request and this is where our e-mail reply scoring method originated from. We also developed a dictionary of favourite users' words this is a dictionary of words that our algorithm select from message contents for each thread of e-mail conversations from each e-mail senders and will note the favourable word that they use when communicating using their email.

We also explore the importance of interrogative words which usually denote a request. These heuristic features form the basis of our solution to e-mail reply prediction. Based on the e-mail subject and content extractions: interrogative words, dictionary of words from senders, most-used phrases, previous e-mail conversations, cc/bcc e-mail addresses, we develop a scoring mechanism for each annotated e-mails and the more score that a mail acquires the more apparent such e-mail needs reply.

All e-mails have same scores zeros at the beginning of the analysis. Negative score is possible. Also each email has been annotated with these properties:

- Definitely need reply - 1,
- Definitely need no reply - 0

If any e-mail has both "definitely need reply" and "definitely need no reply" then these properties delete each other but this case is rare.

The term "definitely need reply" status is given if our algorithm detect phrases such as "please reply soon" and "definitely need no reply" status is given when we found phrases such as "do not reply" or address such as noreply@domain.com.

Our scoring system changes score allocation to each email before making a decision if they need a reply or not and also, if it founds interrogative words or questions or questions mark (s) in email messages, it increases or decreases the score. The other features that we investigated are:

Dictionary of words: If e-mail has many words that interesting to user then increase score, dictionary of favorite users words need. Also algorithm dictionary keeps senders' most used words.

E-mail domains: If senders name is from ".com, .ac.uk, .edu," then decrease/increase score, it is big organization e-mailing

Communications from sender: If communication with sender was earlier then increases score ("Re:"-letters) (user sent emails analysis need).

Interrogative words: Can, Could, Will, When, Where, How, Who is? Who? Which What? etc.

Previous e-mail conversation: If sender send emails earlier and that was not answered then decrease score (user sent emails analysis need).

Email fields: When there is one or more email addresses in "CC/BCC" field then increase score.

Attachments: if big attachment in the e-mail then increase scores (photo or interesting pdf-article from friend).

MACHINE LEARNING TECHNIQUES

E-mail messages in mail boxes could become large amount of data that could have hidden correlations and may be hard to find specific e-mails 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 e-mail address (domain from where this e-mail 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;
- Favorite user dictionary;
- Develop a scoring mechanism etc

Our technique is capable of learning e-mail 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.

Scoring method

Our approach analysed the feature of e-mails namely; phrases, interrogative words, questions and question mark, attachments, early communications of senders and many other aforementioned features in section 3.2 above and our algorithm prediction system (APS) performs unsupervised scoring methods using weighting measures (Salton et al., 1975). All new e-mails have number - score. Then more score then more email need reply. We calculate the weighting scores on the features of the email by implementing a method called “the inner product” with its elements. We collect n numbers of emails using this function below:

$$S_{q,e} = \sum_{t \in T_{q,e}} \left(w_{q,t} . w_{e,t} \right)$$

Here, $w_{e,t}$ is the email-term weight while query-term weight is denoted by $w_{q,t}$ and we also denote these various set:

- The set *E* of e-mails;
- For each term *t*, the set *Et* of emails containing *t*;
- The set T distinct terms in the database and
- The set T_e of distinct terms in e-mails e, and similarly T_q for queries and $T_{q,e} = T_q \cap T_e$

The terms are the features extracted to determine the e-mail prediction namely: phrases, interrogative words, question marks, attachments and many more. When the formula above is applied, the average weighting score is calculated for each email and if it is above the set threshold, then that mail will be categorized as need reply or do not need reply (need no reply) as given relevant item is retrievable without retrieving number of irrelevant items.

Our predictor assigns a weight score to any question (s), question mark (s) found in email subject as well as contents of the mail. For example: A question in the subject has a weight score of 3 point of value and a weight score of 2 in the body of the email message. Do note that a question is a sentence that ends with the sign "?" and start with an interrogation pattern like: "where", "when", etc. Also, a score of 1 is assigned to the following sample features: "if communication with sender was earlier (“Re:”-letters)", emails from specific domain (.ac.uk, .edu), phrases such as “please reply soon”, if there is an e-mail address in cc or bcc, all these are assigned a score of 1. The prediction analysts concluded that the maximum weight score that could be assigned to every email is 10 and choose 7 as the threshold weighting score that a mail must attain before it could be grouped as “need reply- 1” and any email that does not measure up to the threshold will be re-examined and if other factors have been re-assessed and could not meet up with the threshold at the second attempt, then it will be grouped as “do not need reply- 0.

E-mail prediction methods (EPM)

E-mail space is a function of the manner in which terms and term weights are assigned to the various e-mails with an optimum e-mail space configuration that provides an effective performance. If nothing is known about the e-mails under consideration, it suggests that ideal email space is one where emails are jointly relevant to certain user queries and such mails are predicted together ensuring that they will be retrievable jointly in response to the corresponding queries.

Inner product space (Salton et al., 1975) is a vector space of arbitrary (possibly infinite) dimension with additional structure, which, among other things, enables generalization of concepts from two or three-dimensional Euclidean geometry. The additional structure associate to each pair of vectors in the space is called the inner product (also called a scalar product) of the vectors as shown in the formula below:

a = [a1, a2, an] and b = [b1, b2, … , bn] is defined as:

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n$$

For example, the dot product of two three-dimensional vectors
[1, 3, −5] and [4, −2, −1] is

$$[1 \quad 3 \quad -5] \cdot [4 \quad -2 \quad -1] = (1)(4) + (3)(-2) + (-5)(-1) = 3.$$

For two complex vectors the dot product is defined as

$$a \bullet b = \sum a_i \overline{b}_i$$

Where is the complex conjugate of bi. The absolute avoids two weights cancel each other and that enables us to avoid negative weight measures and correct errors in the weighting system. The complex conjugate of a complex number is given by changing the sign of the imaginary part. Thus, the conjugate of the complex number;

Reply prediction algorithm

```
1. Define X as the number of matching needed to mark
   the message needs reply
2. Define Count as the number of matching =0
3. If CC or BCC containse-mail addresses then
      a. Count= Count+1
4. create a rule that
      a. If the contents contains some of these
         words
            i. Count= Count+1
      b. must, should, what about, meeting ,priority,
            i. Count= Count +1
      c. Dear, hello, hi
            i. Count= Count +1
      d. Multiple of "?"
            i. Count = Count +1
      e. Dates or months names
            i. Count = Count+1
      f. AM,PM
            i. Count= Count+1
5. if(Count> X)
      a. then mail need reply
      b. Else
      c. mail doesn't need reply
```

Figure 3. Algorithm prediction System

$$z = a + ib$$

(Where a and b are real numbers) is

$$\overline{z} = a - ib.$$

The complex conjugate is also very commonly denoted by z *. Here is chosen to avoid confusion with the notation for the conjugate transpose of a matrix (which can be thought of as a generalization of complex conjugation). Notice that if a complex number is treated as a matrix, the notations are identical. For example,

$$\overline{(3-2i)} = 3 + 2i$$
$$\overline{7} = 7$$
$$\overline{i} = -i.$$

One usually thinks of complex numbers as points in a plane with a Cartesian coordinate system. The x-axis contains the real numbers and the y-axis contains the multiples of i. In this view, complex conjugation corresponds to reflection at the x-axis.

In order to measure an angle θ, a circular arc centred at the vertex of the angle is drawn, e.g. with a pair of compasses. The length of the arc s is then divided by the radius of the circle r, and possibly multiplied by a scaling constant k (which depends on the units of measurement that are chosen):

$$\theta = \frac{s}{r}(k).$$

The value of θ thus defined is independent of the size of the circle: if the length of the radius is changed then the arc length changes in the same proportion, so the ratio s/r is unaltered. In many geometrical situations, angles that differ by an exact multiple of a full circle are effectively equivalent (it makes no difference how many times a line is rotated through a full circle because it always ends up in the same place). However, this is not always the case. For example, when tracing a curve such as a spiral using polar coordinates, an extra full turn gives rise to a quite different point on the curve as explained by Sidorov et al (2001). Since our annotated e-mails from Enron corpus are treated like a bulk of dataset, we used term weighting with unsupervised techniques with our approach of heuristic techniques to provide a well organised and prioritized email prediction system.

Algorithm prediction systems (APS)

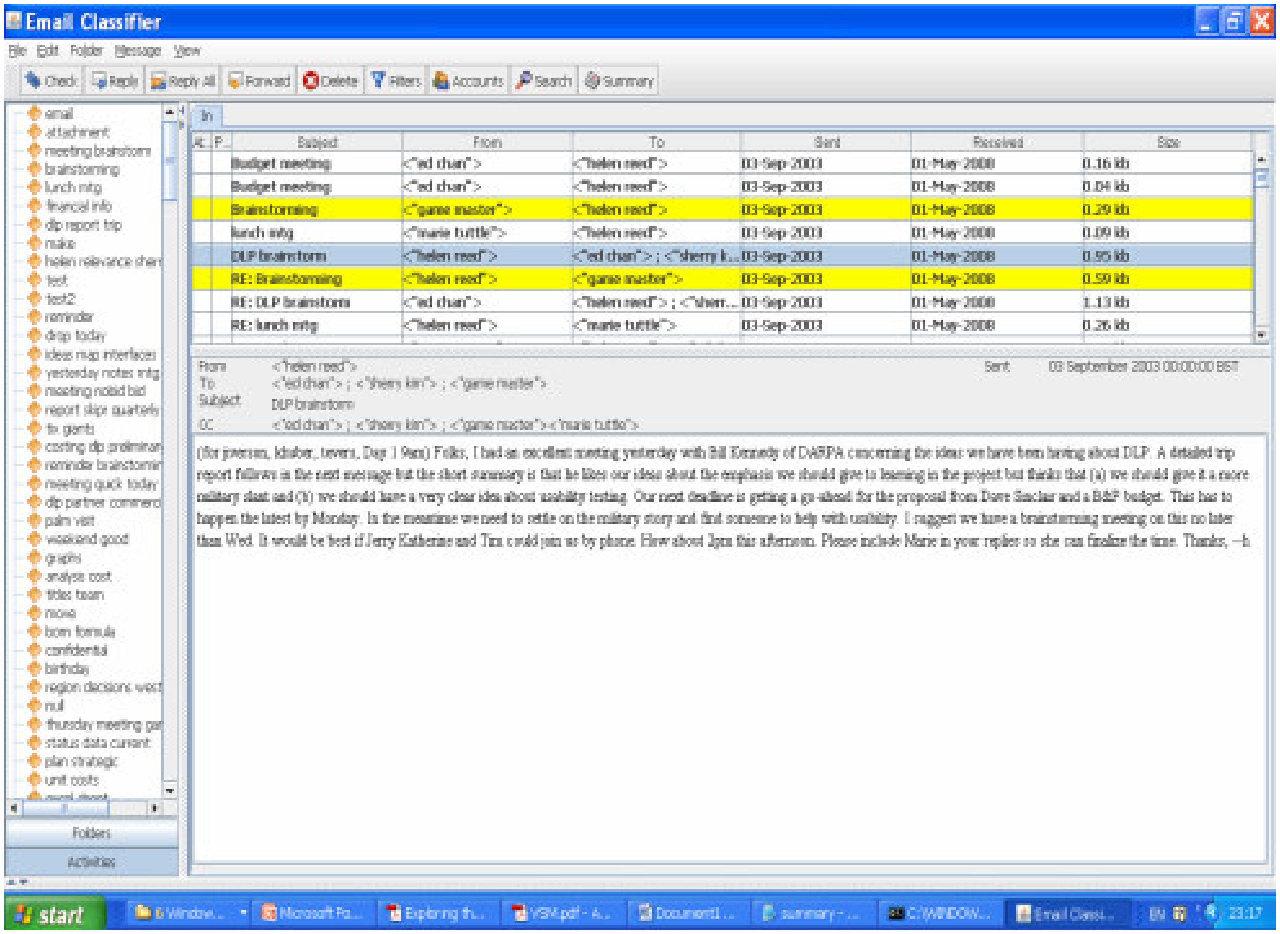
Algorithm prediction system uses a heuristics-based approach with embedded favourite dictionary of word and phrases, with weighting measures. The assumption is that if interrogative words, questions, questions mark(s), phrases such as do reply, when will you, if date and time are found in email messages, such a mail is important and will be assigned some score. The algorithm is shown in Figure 3.

Algorithm prediction system (APS) for email management is a new unsupervised machine learning techniques that is implemented. APS described above uses a precision and recall to evaluate this new technique in comparison with gold- human participant

DATASET SETUP

We collected over 6000 e-mail conversations from the Enron email dataset (Bryan and Yiming, 2004) as the test bed and had 50 human reviewers to review the e-mail prediction system. Notice that having such a gold standard may also be used to verify our assumptions and algorithm. We annotated 6000 emails to determine the original class with numeric values: need reply- 1 and need no reply- 0. We then used human annotated emails as the gold standard to compare our algorithm result with result of human participants.

The 50 human prediction analysts reviewed those 6000 selected email conversations. All the analysts were undergraduate and graduate in university of Portsmouth. Their discipline covered various areas including Science



At.	P...	Subject	From	To	Sent	Received	Size
		Budget meeting	<"ed chan">	<"helen reed">	03-Sep-2003	01-May-2008	0.16 kb
		Budget meeting	<"ed chan">	<"helen reed">	03-Sep-2003	01-May-2008	0.04 kb
		Brainstorming	<"game master">	<"helen reed">	03-Sep-2003	01-May-2008	0.29 kb
		lunch mtg	<"marie tuttle">	<"helen reed">	03-Sep-2003	01-May-2008	0.09 kb
		DLP brainstorm	<"helen reed">	<"ed chan"> ; <"sherry k...	03-Sep-2003	01-May-2008	0.95 kb
		RE: Brainstorming	<"helen reed">	<"game master">	03-Sep-2003	01-May-2008	0.59 kb
		RE: DLP brainstorm	<"ed chan">	<"helen reed"> ; <"sherr...	03-Sep-2003	01-May-2008	1.13 kb
		RE: lunch mtg	<"helen reed">	<"marie tuttle">	03-Sep-2003	01-May-2008	0.26 kb

From <"helen reed"> Sent 03 September 2003 00:00:00 BST
To <"ed chan"> ; <"sherry kim"> ; <"game master">
Subject DLP brainstorm
CC <"ed chan"> ; <"sherry kim"> ; <"game master"><"marie tuttle">

(for jeverson, ldtaber, tevers, Day 1 9am) Folks, I had an excellent meeting yesterday with Bill Kennedy of DARPA concerning the ideas we have been having about DLP. A detailed trip report follows in the next message but the short summary is that he likes our ideas about the emphasis we should give to learning in the project but thinks that (a) we should give it a more military slant and (b) we should have a very clear idea about usability testing. Our next deadline is getting a go-ahead for the proposal from Dave Sinclair and a B&P budget. This has to happen the latest by Monday. In the meantime we need to settle on the military story and find someone to help with usability. I suggest we have a brainstorming meeting on this no later than Wed. It would be best if Jerry Katherine and Tim could join us by phone. How about 2pm this afternoon. Please include Marie in your replies so she can finalize the time. Thanks, --h

Figure 4. A sample reply prediction system.

and Engineering, Arts, Education, Law, Business and IT. Since many emails in the Enron dataset (Bryan and Yiming, 2004) relate to business, IT and law issues, the variety of the human prediction analysts, especially those with business and legal background are of asset to this user study. Each prediction analysts reviewed 120 distinct email conversations in 3 h.

For each email features extracted as described above human prediction analysts (hpa) retain the highest weighting score. Analysts then choose threshold and emails found above this threshold to be categorized as need reply and any emails that does not reach up to the threshold will be categorized as do not need reply. Thus, our expectation that human-annotated email prediction will show great variation was borne out, we discuss these differences further in section 5.

EVALUATIONS AND RESULTS

In order to compare different approaches of email reply prediction, a gold standard is needed. In practice, for comparing extractive predictor, we tested our algorithm performance with 6000 annotated emails from the human participant to:

- Need reply
- Need no reply

We tested our algorithm with the embedded similarity measure approach on the 6000 email datasets. To measure the quality and goodness of the email prediction, gold standards are used as references. It is noticed that our unsupervised machine learning approach achieved 98% accuracy in comparison to the gold standard. Our sample graphical prediction client output is shown in Figure 4.

This section describes experiments using APS system to automatically induced email features classifiers, using the features described in Section 4. Like many learning programmes, APS takes emails as input and the classes to be learned, a set of features names and possible values and training data specifying the class and feature values for each training example. In our case, the training

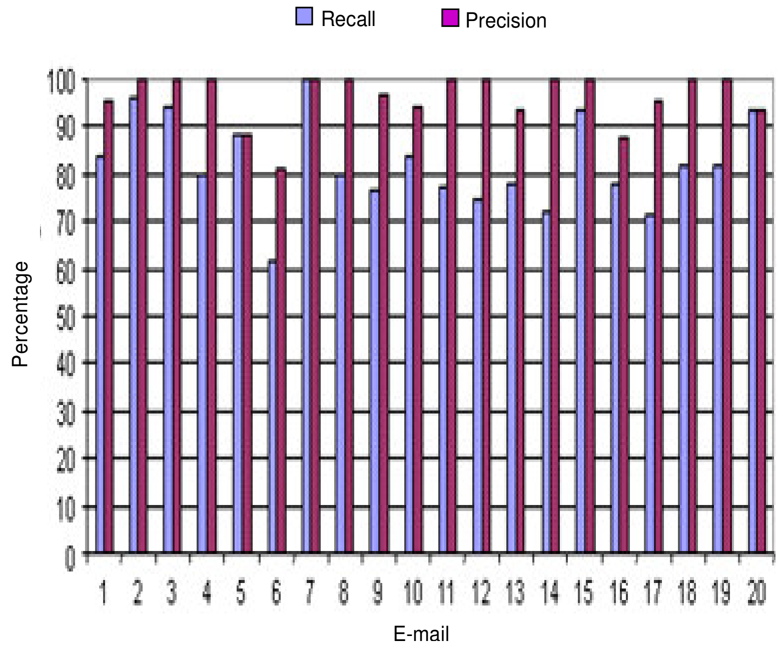


Figure 5. Evaluation result.

Precision and recall (compared to gold standard)

Correct predicted group	Total Predicted Group Found	Total E-mails	Precision (%)	Recall (%)
309	316	382	98.0	80.9

examples are the Enron email datasets. APS outputs a classification model for predicting the class (that is, need reply- 1, need no reply- 0). We obtained the results presented here using precision and recall. In this paper, we evaluated APS system based on weighting measures, and human judgments. We show results of 6000 annotated emails and different feature set in Figure 5.

We evaluated our email reply prediction system on over 6000 email Enron datasets from over 120 email boxes owned by 200 people from Enron Corpus using precision and recall. We evaluate our proposed e-mail prediction system against human email predictions. Human participant detected replies by matching e-mail features: interrogative words in email contents, phrases (reply soon, need your help), previous email conversations, attachments, interesting words as chosen by email user from dictionary, question mark (s) in emails and reference fields of a message with message Id from original message. 50 participants were involved. Figure 5 shows more evaluation results.

These second participants were separated into two groups and were given 1500 emails to analyse, annotate and predict the mails that require a reply. Group 1 confirmed that out of 1500 e-mails, only approximately 668 require a reply while group 2 confirmed that approximately 665 require a reply. With the human participants, the average estimation of mails that need reply is 667 which could mean that human judgement in this case could be term as 100% accurate. Our proposed email reply prediction system estimated that approximately 658 require a reply and is approximately 98% accurate.

We also evaluate our algorithm prediction system using precision and recall as the measurement of evaluation for our system:

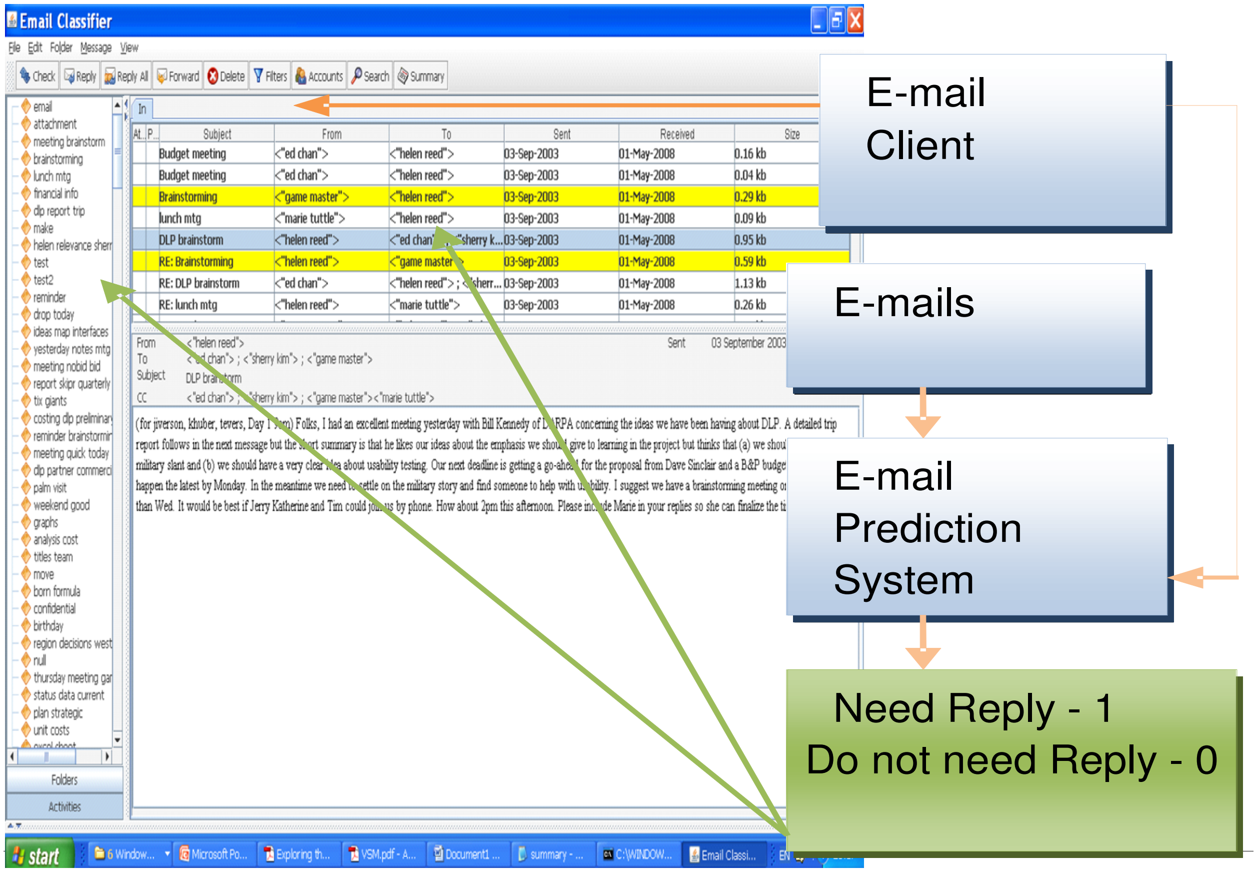
- For 1500 emails, compute the recall and precision where a correct predicted group is found.
- Given our prediction system whose input are email messages, and whose outputs are need reply and need no reply. The recall and precision is computed as:

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

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

We evaluate our prediction algorithm's performance by comparing performance of human participants our new proposed prediction system. Figure 5 shows detail results.

Figure 5 also shows a second evaluation test results with the accuracy of our precision and recall evaluation on 1500 e-mails but we show only results of 382 e-mail datasets because of limited space. The email prediction system relies on a simple algorithm but it is very complex to implement. Yet it appears to work better than other existing approaches. Since our algorithm prediction sys-



At..	P...	Subject	From	To	Sent	Received	Size
		Budget meeting	<"ed chan">	<"helen reed">	03-Sep-2003	01-May-2008	0.16 kb
		Budget meeting	<"ed chan">	<"helen reed">	03-Sep-2003	01-May-2008	0.04 kb
		Brainstorming	<"game master">	<"helen reed">	03-Sep-2003	01-May-2008	0.29 kb
		lunch mtg	<"marie tuttle">	<"helen reed">	03-Sep-2003	01-May-2008	0.09 kb
		DLP brainstorm	<"helen reed">	<"ed chan"> ; <"sherry k...	03-Sep-2003	01-May-2008	0.95 kb
		RE: Brainstorming	<"helen reed">	<"game master">	03-Sep-2003	01-May-2008	0.59 kb
		RE: DLP brainstorm	<"ed chan">	<"helen reed"> ; <"sherr...	03-Sep-2003	01-May-2008	1.13 kb
		RE: lunch mtg	<"helen reed">	<"marie tuttle">	03-Sep-2003	01-May-2008	0.26 kb

From <"helen reed"> Sent 03 September 2003
To <"ed chan"> ; <"sherry kim"> ; <"game master">
Subject DLP brainstorm
CC <"ed chan"> ; <"sherry kim"> ; <"game master"><"marie tuttle">

(for jiverson, khuber, tevers, Day 1 9am) Folks, I had an excellent meeting yesterday with Bill Kennedy of DARPA concerning the ideas we have been having about DLP. A detailed trip
report follows in the next message but the short summary is that he likes our ideas about the emphasis we should give to learning in the project but thinks that (a) we shoul
military slant and (b) we should have a very clear idea about usability testing. Our next deadline is getting a go-ahead for the proposal from Dave Sinclair and a B&P budge
happen the latest by Monday. In the meantime we need to settle on the military story and find someone to help with usability. I suggest we have a brainstorming meeting o
than Wed. It would be best if Jerry Katherine and Tim could join us by phone. How about 2pm this afternoon. Please include Marie in your replies so she can finalize the ti

E-mail Client

E-mails

E-mail Prediction System

Need Reply - 1
Do not need Reply - 0

Figure 6. Graphical output.

tem is built on weighting measures, we believe this has helped the accuracy of our result. Figure 6 shows the prediction client.

E-mail messages are passed unto our prediction system as shown in Figure 6 and our prediction algorithm extracts the features that makes up the prediction decision from each email messages and intelligently determine the mail that needs reply showing the output of need reply as numeric value 1 and need no reply as numeric value 0. In Figure 6, the yellow flag indicates emails that require attention-needs reply and the rest of the e-mail in the mail box remain as normal mails.

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 any reply. We analyse the features of emails and study email conversation structure, users' favourite dictionary which we maintain that this area of research has not been sufficiently investigated in previous research on email reply prediction. We build a novel structure: Algorithm prediction system (APS), interrogative words, mails from specific domains and many more.

Our future plan includes improving the algorithm prediction system with more sophisticated linguistic analysis. And implementing security into the email prediction system. 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 requires a lot of efforts and in conclusion, explore how to combine our prediction techniques with several machine learning algorithms.

REFERENCES

Whittaker S, Sidner C 1996. 'Email overload: exploring personal information management of email', In The Proceedings of the ACM CHI 96 Human Factors in Computing systems Conference, ACM Press, pp.276-283.

Bradley JR, Thad S (1996). 'the remembrance agent: A continuously running information retrieval system', In the Proceedings of the First International Conference on Practical Applications of Intelligent Agents and Multi-Agent Technology (PAAM'96), London, pp.486–495.

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

Salton G, Wong A., Yang CS (1975). *A vector- space model for automatic indexing. Communications of the ACM*. 18(11) p613-620

Laura D, Gina V, Cadiz JJ (2003). Marked for deletion: An analysis of email data', In the Conference on Human Factors in Computing Systems CHI '03 extended abstracts on Human factors in computing systems **(***CHI 2003)*, ACM Press, pp.924-925.

Ducheneaut N, Belloti V (2001). Email as habitat: An exploration of embedded personal information management. ACM Interactions. 8(5) pp.30-38.

Tyler JR, Tang JC (2003). When can i expect an email response? A study of rhythms in email usage. In Proceedings of the Eighth Conference on European Conference on Computer Supported Cooperative Work (Helsinki, Finland, September 14 - 18, 2003). Kuutti K, Karsten EH, Fitzpatrick G, Dourish P, Schmidt K, Eds. ECSCW. Kluwer Academic Publishers, Norwell, MA, 239-258.

Mackay W (1998). Diversity in the use of electronic mail: A preliminary inquiry. *ACM Transactions on Office Information Systems.* ACM Press. New York, NY, USA. 6(4) pp.380-397.

Sproull L, Kiesler S (1991). *Connections: New ways of working in the networked organization*. MIT Press: Cambridge, MA, USA.

Mary S (1985). 'The impact of electronic mail on managerial and organizational communications', *In the Proceedings of ACM SIGOIS and IEEECS TCOA Conference on Office Information Systems.* ACM Press. New York, NY, USA. pp.96-109.

Kraut RE, Paul A (1997). Media use in a global corporation: Electronic mail and organizational knowledge, In Culture of the Internet. Lawrence Erlbaum, Associates, Mahwah, NJ, USA.pp.323-342.

Bryan K, Yiming Y (2004). The Enron corpus: A new dataset for email classification research. In European Conference on Machine Learning.

Sidorov L.A. (2001) . *Angle*. Encyclopaedia of Mathematics.