A GENERALIZED EMAIL CLASSIFICATION SYSTEM FOR WORKFLOW ANALYSIS

Piyanuch Chaipornkaew, Dhurakij Pundit University
Takorn Prexawanprasut, Dhurakij Pundit University
Chia-Lin Chang, National Chung Hsing University
Michael McAleer, Asia University, University of Sydney, Erasmus University Rotterdam, Complutense University of Madrid, Yokohama National University

ABSTRACT

One of the most powerful internet communication channels is email. As employees and their clients communicate primarily via email, much crucial business data is conveyed via email content. Where businesses are understandably concerned, they need a sophisticated workflow management system to manage their transactions. A workflow management system should also be able to classify any incoming emails into suitable categories. Previous research has implemented a system to categorize emails based on the words found in email messages. Two parameters affected the accuracy of the program, namely the number of words in a database compared with sample emails and an acceptable percentage for classifying emails. As the volume of email has become larger and more sophisticated, this research classifies email messages into a larger number of categories and changes a parameter that affects the accuracy of the program. The first parameter, namely the number of words in a database compared with sample emails, remains unchanged, while the second parameter is changed from an acceptable percentage to the number of matching words. The empirical results suggest that the number of words in a database compared with sample emails is 11 and the number of matching words to categorize emails is 7. When these settings are applied to categorize 12,465 emails, the accuracy of this experiment is approximately 65.3%. The optimal number of words that yields high accuracy levels lies between 11 and 13, while the number of matching words lies between 6 and 8.

Keywords: Email, Business Data, Workflow Management System, Business Transactions.

JEL: J24, O31, O32, O33.

INTRODUCTION

Information and communication technology has been developed significantly in recent years. The technology eliminates the wall of distance and connects people more closely than ever. The technology also supports many businesses to gain competitive advantages. Owing to this technology, large numbers of organizations are able to operate their business at lower costs and with a higher competitive advantage. As a result, many organizations attempt to acquire this on-time and accurate information. One of the most powerful tools in business is email, which is a fundamental and indispensable communication channel for every organization in the modern age.

In recent decades, the number of start-up companies has increased dramatically. Two of the authors have participated in three start-up companies related to the import/export sector. These new start-ups established their own businesses by separating themselves from their former companies. After the initial study, it was found that start-up companies needed to manage a large number of daily documents/emails because start-up businesses contacted their customers and employees primarily via email. The employees also used these emails, which were stored in the mail server, as a database. For example, when employees wanted to find specific data, emails were the first place for seeking information.

In the first stage of starting their businesses, the number of emails was not large. However, when the scale of business expanded, the number of emails increased. The business owners needed applications to manage their company activities, a problem that could be solved primarily by software applications, such as the workflow management system. However, the cost of this software is rather high and may not be appropriate for start-up companies, so that alternative approaches to solve the problem were needed.

For the initial investigation, 12,465 of emails were selected from the three start-up companies because they were written in English. As the employees in the selected companies wrote emails in two languages, namely English and Thai, only emails that were written in English were taken into consideration as the sentences in English are easier to separate into words than corresponding emails in Thai. By investigating some of these emails, some keywords specified the type of work, such as sales, transportation, billing or shipping, which can be used as initial guidelines to conduct the classification models.

The purpose of this paper is to define the categories of email and extract business data for a workflow management system.

The remainder of the paper is as follows. Section 2 provides a literature review, Section 3 describes the materials and methods, Section 4 presents the data analysis, Section 5 illustrates the results and discussion and Section 6 provides some concluding comments.

LITERATURE REVIEW

There is much research that mentions the clustering and classification of email content and many objectives to conduct research for email classification problem, such as: Distinguishing between personal and machine-generated email (Mihajlo, Halawi, Karnin and Maarek, 2014), classifying emails for contact centres (Nenkova and Bagga, 2003); classifying emails for automated service handling (Taliby, Dean, Milner and Smith, 2006); and classifying emails for social network analysis (Yelupula and Ramaswamy, 2008). As regards classification techniques, there are also many methods applied to email classification, such as mining-based approaches (Aery and Chakravarthy, 2004), supervised learning algorithms (Tam, Ferreira and Lourenco, 2012), co-training technique (Kiritchenko and Matwin, 2001; Kiritchenko and Matwin, 2011), co-training with a Single Natural Feature Set (Chan, Koprinska and Poon, 2004) and regression-based approaches (Yoo, Yang and Carbonell, 2011).

One of the interesting topics is by Alsmadia and Alhamib (2015). The authors illustrate that the best algorithm to perform email clustering and classification is NGram. Their sets of emails were in the form of a large text collection, which fits with the NGram algorithm and the algorithm best fits the bi-language text. They conducted an experiment based on emails in both English and Arabic. The major challenge of their future work was that email servers or applications should include different types of pre-defined folders. The general pre-defined folders could be mailbox, sent or trash, among others. Moreover, email servers or applications could allow users to add new folders for specific purposes, based on their NGram algorithm.

Further research on email classification is by Katakis, Tsoumakas and Vlahavas (2006). They state that Machine Learning and Data Mining could be used as tools to automate email managing tasks, which could be far superior to other conventional solutions. They discuss the particularity of email content and what special treatment it requires. In addition, there are some interesting email mining applications, like mail categorization, summarization, automatic answering and spam filtering. In their experiments, they created an application to classify email based on several techniques, such as the Naïve Bayes Classifier and Support Vector Machines.

Ayodele, Khusainov and Ndzi (2007) present the design and implementation of a system to group and summarize email messages. Their system considers the subject and content of email messages to classify emails based on user activities and produces summaries of each incoming message with an unsupervised learning approach. They claim that their framework could solve the problems of email overload, congestion, difficulties in prioritizing and difficulties in finding previously archived messages in the email server.

Another interesting topic is email grouping and summarization. Ayodele, Zhou and Khusainov (2009), present the design and implementation of an application to categorize and summarize email content. Their system extracts the subject and content of email messages for classification based on user activities to auto-generate a summary of each incoming message. They state that their framework could solve problems such as email overload, difficulties in prioritizing and email congestion. Their framework also performs successful processing of new incoming messages.

Another interesting concept is automated email activity management, as in Kushmerick and Lau (2005), who develop email applications that provide high-level support for structured activities in e-commerce. They define formal activities as finite-state automata, which correspond to the status of the process and where transitions represent messages sent between participants. They propose several unsupervised machine learning algorithms and evaluate a collection of e-commerce emails.

Schuff, Turetken, D'Arcy and Croson (2007) also discuss email classification. They implement effective e-mail management tools, which treat messages as useful information. This tool could economize on scarce cognitive resources at the expense of relatively cheap additional CPU power, disk capacity and network bandwidth. In addition, they claim that their application provides automatic filtering, clustering and a new user interface. Their system employs a large number of emails as an effective knowledge management tool, rather than as a source of information overload.

Email classification is discussed in Prexawantprasut and Chaipornkaew (2017). The research classifies email into four categories, namely sales, shipping, billing and transportation. Two parameters are applied for the classification system, namely the number of words in a database compared with the sample emails and an acceptable percentage to classify emails. The accuracy of classification is determined to be approximately 73.6%.

Chaipornkaew, Prexawanprasut and McAleer (2017) discuss email extraction for workflow management system. In order to extract data, there are four criteria which are applied. Fifteen cases of alternative criteria to extract data are analysed. The results show that when criteria numbers 2 and 4 are considered, email extraction accuracy is at the highest level. However, when the highest accuracy level occurs, the number of blanks fields is also high. According to user requirements, the number of blank fields should be at a low level. Therefore, the paper suggests that all four criteria should be considered to provide both an acceptable percentage of blank fields and also accuracy level.

MATERIALS AND METHODS

The paper is planned in two phases, as shown in Figure 1. First, 1260 emails are selected randomly from the server to be used as training data for the system. These emails are then classified manually by employees into seven categories, namely (1) Sales, (2) Agent, (3) Shipping, (4) Customs, (5) Billing, (6) Packing and Moving and (7) Insurance. The sentences in emails are separated into words, which are counted, as shown in Figure 2a, 2b. These results are stored in the database, which is applied for email classification rules.

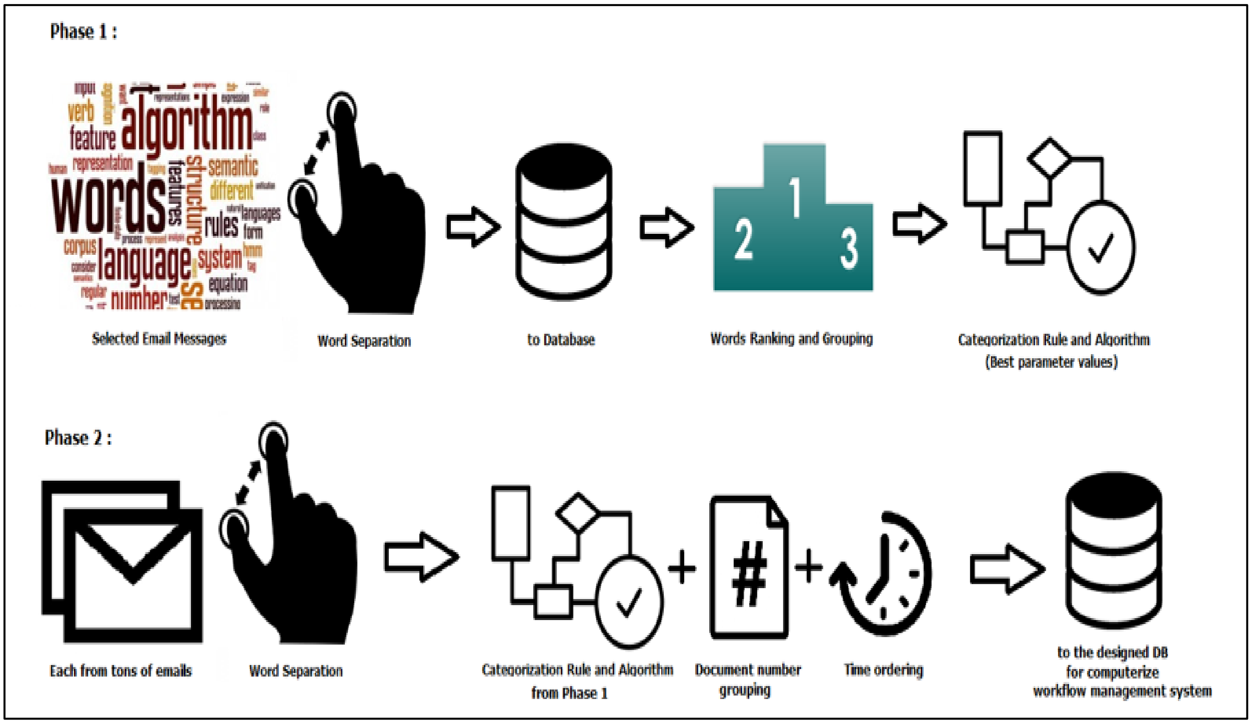

FIGURE 1
TWO PHASES OF THE EMAIL CLASSIFICATION SYSTEM

In order to test the defining rules, a further 12,465 emails are selected from the server. When these rules are accepted, the rest of the emails in the server are processed by the program. After the classification is processed, all emails are assigned to suitable categories and then all the data are prepared for the second phase of the email classification system.

	about	Above	according	acknowledge	adding	advise	after	again	ahead
Sales	4	1	2	3	1	1	0	0	2
Agent	3	0	2	2	1	1	0	0	3
Shipping	1	2	1	3	2	5	1	0	4
Customs	0	0	2	1	4	2	0	1	1
Billing	2	0	1	1	2	1	1	0	2
Packing and Moving	2	3	2	1	1	2	0	1	2
Insurance	1	0	1	0	2	3	1	0	1

	all	already	and	any	anytime	apply	around	arrange	Arrival	attached
Sales	2	1	2	1	2	0	3	1	5	8
Agent	1	3	1	0	5	2	1	1	1	6
Shipping	2	2	1	0	1	2	2	0	1	2
Customs	1	2	0	1	3	3	2	0	3	7
Billing	0	1	0	2	2	4	2	6	2	5
Packing and Moving	3	0	1	3	1	4	4	2	4	4
Insurance	2	1	2	1	2	1	1	7	1	3

FIGURE 2
EXAMPLE OF RESULTS FROM THE WORD SEPARATION PROCESS

The second phase is to extract the classified emails, which are processed from the first phase. As in investigating the selected emails, there are key characteristics which can be represented as relationships. For example, the document number could be a key characteristic to define the relationships among the email messages. The program first reorders emails based on time in each category, then extracts data based on their characteristics. The final stage is to create a workflow management system from the extracted data.

DATA ANALYSIS

The first stage is to export all emails from the email server and format them in a text file, which is then imported to the program. The program first separates words in a text file. As the selected emails are in English, the algorithm to separate the words is the use of spaces. The words from the separation process are counted and stored in a database. The database stores all results which are all words and their frequencies as shown in Table 1.

Table 1
TOP 15 WORDS IN EMAILS IN 7 CATEGORIES

Sales		Agent		Shipping		Customs	
Word	**Frequency**	**Word**	**Frequency**	**Word**	**Frequency**	**Word**	**Frequency**
agent	112	#NAME of CUS	188	shipment	167	tax	109
volume	91	arrange	165	scheduled	112	standard	87
#NAME of CUS	88	ETA	150	ETA	102	customs	74
product	72	delivery	112	#Date format	89	clear	52
shipment	60	#NAME of CITY	94	ship	82	#Date format	43
#NAME of CITY	58	import	86	D/O	80	scheduled	42

process	55	items	81	shipper	65	#NAME of port	38
confirm	52	#NAME of PORT	75	#NAME of CUS	55	shipment	33
week	48	warehouse	53	#NAME of CITY	42	departed	28
#Date format	31	service	50	confirm	40	#NAME of PORT	25
D/O	28	update	48	HBL	35	fare	23
packing list	25	port	39	BL	32	transaction	22
#NAME of PORT	22	shipping	31	port	21	notification	18
attach	18	scheduled	21	#NAME of PORT	19	standard	16
request	16	#Date format	12	request	17	arrived	15

Billing		**Packing and Moving**		**Insurance**	
Word	**Frequency**	**Word**	**Frequency**	**Word**	**Frequency**
consignee	125	loading	108	policy	78
shipper	111	destination	75	dividend	62
document	94	package	71	product	55
revise	89	carrier	60	fair	53
#NAME of CUS	84	loader	65	#NAME of CUS	42
scheduled	74	#Date format	55	accident	41
departed	62	co-loader	48	rate	28
service	50	departed	40	title	24
#Date format	48	ETD	34	revenue	22
arrived	42	arrived	33	package	22
#NAME of PORT	38	scheduled	33	#Date format	21
shipment	31	#NAME of PORT	32	arrived	13
notice	25	shipment	28	departed	13
booking	22	worker	24	loss	11
approval	18	condition	15	value	10

The research classifies 12,465 emails into seven categories, namely (1) Sales, (2) Agent, (3) Shipping, (4) Customs, (5) Billing, (6) Packing and Moving and (7) Insurance. The mechanism is implemented based on the words found in emails compared with the words in the database for each category. Two parameters are considered in this experiment. The first parameter is the number of words in the database. For example, in order to gain greater accuracy in the classification, we need to determine whether the first 3 or 5 words in the database should be considered. The second parameter is the number of matching words that provides the highest accuracy to determine the category of email.

Table 2
GROUPING RESULTS BASED ON TOP 5 WORDS AND 5 ACCEPTABLE NUMBER OF MATCHING WORDS

No. of Emails	Number of Matching Words							Grouping result
	Sales	Agent	Shipping	Customs	Billing	Packing and Moving	Insurance	
1	5	3	0	2	1	1	0	Sales
2	5	0	4	1	1	5	1	Sales or Packing and Moving
3	0	0	1	2	0	1	2	Uncategorized
4	1	0	2	3	5	0	0	Billing
5	1	0	4	0	3	2	5	Insurance

Note: In the case of email no. 2, it falls into either Sales or Packing and Moving category. The research could not conclude whether it should be in the Sales or Packing and Moving group. This issue should be clarified in future research.

According to the data in Table 2, some email could not be classified because the number of matching words is less than the specified criteria. In this case, the second criterion is the first 5 words in a database. In order to obtain better results, these two criteria may need to be refined. As shown in Table 3, the first 10 words in a database are considered instead of the first 5 words.

Table 3
Grouping Results Based on Top 10 Words and 4 Acceptable Number of Matching Words

No. of Emails	Number of Matching Words							Grouping result
	Sales	Agent	Shipping	Customs	Billing	Packing and Moving	Insurance	
1	6	2	4	2	1	1	0	Sales
2	4	0	2	1	1	5	1	Packing and Moving
3	0	4	1	2	0	1	3	Agent
4	1	0	4	4	7	0	0	Billing
5	1	4	8	0	3	2	1	Shipping

The number of matching words is set at 5 in Table 2 and set at 4 in Table 3. As a result, only two groups of output in Tables 2 and 3 are the same. The first difference is the No. 2 group of emails. In Table 2, Email No. 2 could be either Sales or Packing and Moving, but it is concluded to be Packing and moving group in Table 3. The second difference is the No. 3 group of emails, which could not be grouped in Table 2, but could be defined as Agent in Table 3. The third difference is the No. 5 group of emails, which is defined as Insurance group in Table 2, while in Table 3 it is concluded to be shipping.

The empirical data from both Tables 2 and 3 demonstrate that are two main factors that affect the grouping results. The first factor is the number of words in the database to be considered, while the second factor is the number of matching words. Therefore, another 12,465 emails are collected to test the program by changing the criteria for these two factors, with the empirical results shown in Figure 3.

Number	Number of Matching Words																			
of Words	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
5	13.7	27.9	30.6	30.1	29.9															
6	14.4	20	24	32.1	32.5	33.1														
7	18.2	19.5	20.1	23.6	36.8	32.3	23.3													
8	16.2	17.5	18.2	32.5	40.1	25.2	20.1	18.3												
9	15.2	23.6	38.8	42.6	42.1	39.6	35.3	35.1	23.5											
10	19.2	18.2	35.5	48.5	47.8	50.1	38.5	26.5	28.2	20.5										
11	14.6	23.5	28	36.3	45.1	56.8	65.3	52.2	35.1	18.2	10.8									
12	15.2	18.5	20.4	25.7	40.6	58	58.1	60.1	39	20.3	18.8	12.8								
13	18.7	19.6	21.5	26	47.5	63.3	54.9	62.1	41	25.1	19.5	15.1	15.6							
14	14.3	20.1	12.1	15.6	20.7	32.2	34	45.6	28.9	29.4	18.8	12.3	10.2	8.75						
15	14.5	15.7	13.3	12.1	15.5	18.3	20.3	14.5	13.3	12.5	10.8	8.45	8.5	7.85	12.3					
16	12	19	15.7	13.2	12.6	11.9	10.3	18.5	15.2	12.3	14.6	13.8	17.4	12.1	11.5	10.6				
17	6.02	14.6	12.3	11.9	10.6	8.72	14.3	13.2	12.6	11.6	13.2	17.6	15.5	13.2	11.2	9.25	10.8			
18	10.6	12	10.2	8.92	9.95	8.75	12.2	15.3	10.8	18.5	12.2	12	8.16	7.13	10.6	7.75	10.2	9.5		
19	8.12	9.87	10.2	13.9	10.3	13.8	15.5	18	15.9	13.5	13	13.3	14.1	8.54	12.3	8.63	8.72	9.58	9.98	
20	9.5	7.14	12.5	8.12	10.2	11.5	15.5	13.9	12.5	11.6	10.1	13.5	12.2	10.3	5.56	6.23	8.59	6.54	7.63	9.72

FIGURE 3
ACCURACY (%) OF EMAIL CLASSIFICATION

RESULTS AND DISCUSSION

The results shown in Figure 3 illustrate that the accuracy levels change when the number of words in the database and the number of matching words change. The purpose of the paper is to discover suitable parameter values, namely: (1) The number of words in the database to be considered; and (2) the number of matching words. The number of words in a database to be considered is adjusted from 5 to 20, while the numbers of matching words are adjusted from 1 to 20.

According to the results in Figure 3, the highest accuracy level of email classification occurs when the number of words in a database is 11 and the number of matching words is 7. Therefore, these criteria are applied in the program. The program then classified the other 12,465 emails into seven groups, namely: (1) Sales, (2) Agent, (3) Shipping, (4) Customs, (5) Billing, (6) Packing and Moving and (7) Insurance, as shown in Table 4.

Table 4
NUMBER OF EMAILS IN EACH CATEGORY

Sales	Agent	Shipping	Customs	Billing	Packing and Moving	Insurance	Unclassified	Total
1,994	1,623	1,246	1,121	1,371	1,745	872	2,493	**12,465**

According to Table 4, the program could not categorize all the emails because some emails do not meet the acceptable criteria. The program is able to define only 9,972 emails from a total of 12,465 emails, which represents 80% of the total. There are 2,493 emails which could not be categorized in the experiment. In order to improve the program efficiency, other factors could be concerned. One possible factor could be the importance level of each word (the weight of each word) in a database. For example, words that are found most frequently in emails should be placed at a higher level of importance than those that are found less frequently.

When the first phase is completed, all emails are already classified into groups (Sales, Agent, Shipping, Customs, Billing, Packing and Moving and Insurance). The next phase is to analyse the characteristics of the emails. Key characteristics are defined by employees. The program collects these characteristics, which are applied for data extraction. The program reorders the events based on time in each category, as shown in Figure 4.

```
MailNo 4584 anonymous@anonymous - date time of sending
As per the shipper the goods are ready ; here is rate for august departure
ETD 08/06 ( cl 07/31) on Toledo triumph ETA 09/13
ETD 08/13 ( cl 07/31) on Thalassa triumph ETA 09/20
QUOTE REFERENCE
Terms Exw
Port of loading ANTWERPEN
Port of discharge BANGKOK - klong toey terminal
Pick up place SAINGHAIN
Shipping line EVERGREEN
Transit time 41 days via Kaohsiung
Frequency Weekly
Equipment 40' hc
Commodity General cargo
Dangerous goodsyes
Validity 31/08/17

MailNo 4585 - anonymous@anonymous - date time of sending
Thank you for you reponse.
Im waiting you reply with me.
Sincerely your,

MailNo 4593 - anonymous@anonymous - date time of sending
Good day to you. i am checking all details and come back to you quickly
Kind regards

MailNo 4596 - anonymous@anonymous - date time of sending
Good day.
Kindly you contact with shipper for arrance shipment & reques rate ex-work with me.
I looking forward kind to you reply soon.
Thank Best Regard,

MailNo 4599 - anonymous@anonymous - date time of sending
We are pleased to confirm you that your order is ready for collection. Please check with your forwarder to proceed as
follows:
Pick-up Ref: 171055143
Quantity : 19 pallets 120x80x115 cms
Gross weight : 8948 kgs
Hazardous products
Incoterm : ExWorks Sainghin-en-M?lantois
Pick-up address :
LABORATOIRES ANIOS
3330 Rue de Lille
59262 Sainghin-en-M?lantois
FRANCE
Opening hours: 8h00-18h00 (17h on Fridays)
Please find enclosed copy of invoice & DGD.
We thank you to book with me for collection date.
According to European Regulation N?649/2012, any substance or mixture containing didecyldimethylammonium
chlorure (DDAC), with quantities equal or superior to 0,25%, must be submitted to exportation notification procedure
of the importer country (excluding UE28 and some DOM-TOM) starting 01/12/2014.
As a consequence, a RIN Number will be added to our invoices for each concerned product. This Number is intended
to Customs for verification of the procedure application and must appear on EX A.
Best regards,

MailNo 4620 - anonymous@anonymous - date time of sending
Thanks for this new order.
Please find enclosed our confirmation order,
Wish you a good receipt,
Best regards,
```

FIGURE 4
PROGRAM RESULTS AFTER GROUPING, EVENT ORDERING AND INCLUSION OF EMAIL CHARACTERISTICS

The last stage is to extract the specified data based on their characteristics. As the characteristics of data are in many forms, the extracted data can vary substantially. One example of data which are extracted based on Document Number (FWO0018) is shown in Figure 5. According to the results, all the details concerned with Document Number (FWO0018) are well summarized. The data that are extracted will be stored in a database, which will be implemented for a workflow management system.

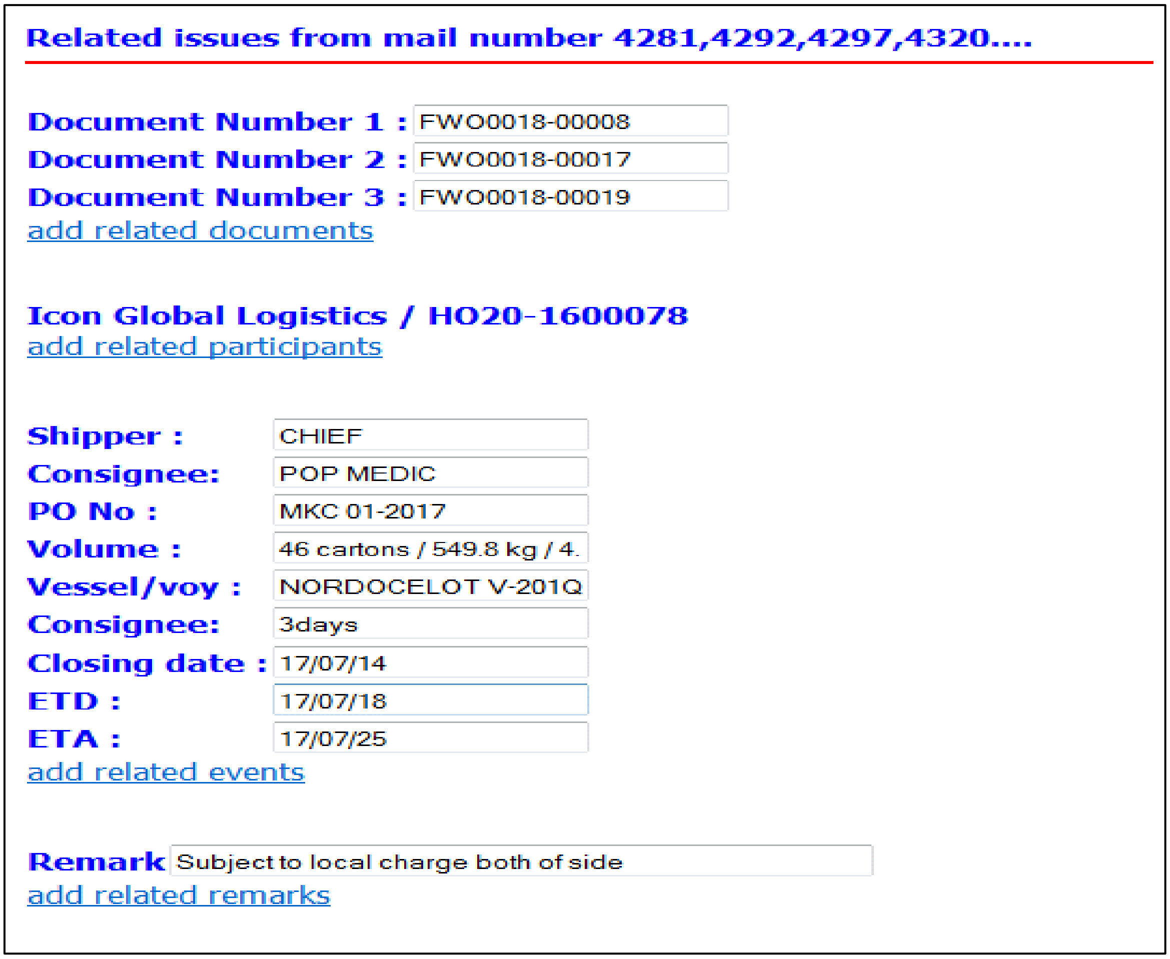

FIGURE 5
EXAMPLE OF EXTRACTED DATA FROM PHASE 2, BASED ON DOCUMENT NUMBER

CONCLUSION

According to the experiments, the accuracy level of email classification depends on two factors, namely the number of words to be considered in a database and the number of matching words. After testing the program with different values for these two factors, the results show that the optimal value for the number of words in a database is 11, while the number of matching words is 7. The results also illustrate that high accuracy levels fall in the range of the number of words lying between 11 and 13, while the range of the number of matching words lies between 6 and 8.

As mentioned earlier, the experiments select all emails in English, so some words need to be neglected. Examples of words which should not be considered are 'and', 'not', 'thanks', 'regards' and 'please'. As these words could be found in most emails, they should not be included in the program. As these words could not be used as criteria to classify email, a more sophisticated program should be developed to ignore these words before processing the email classifications.

In investigating email content, there are specific words that should not be used as criteria in email classification. Examples of these words are FREIGHTLINKS, STARSHIP and HERMESINT'L. As these words are actual customer names, they should be defined as customer names in the database and are excluded from the criteria for email classification in the first phase. However, these specific data are the key characteristics for the second phase of the research. The data with their characteristics are applied to extract data, which are used for the workflow management system.

The generalized email classification system for workflow analysis has been shown to work well in the experiments, with a high degree of accuracy.

ACKNOWLEDGEMENT

The authors would like to thank the Executive Vice-President of the Finish International Freight Co. Ltd., as well as two anonymous companies which cannot be mentioned as a result of confidentiality. The companies provided the useful information to conduct this research. Thanks also to Khun Natthicha Phonjan and Khun Sariporn Plipon, who assisted in manually classifying the emails. It is appreciated that the business data provided by the three selected businesses are sensitive and will not be disclosed or used for any purpose other than the research for the paper.

REFERENCES

Aery, M. & Chakravarthy, S. (2004). EMailShift: Mining-based approaches to email classification. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 580-581.

Alsmadia, I. & Alhamib, I. (2015). Clustering and classification of email contents. *Journal of King Saud University-Computer and Information Sciences, 27*(1), 46-57.

Ayodele, T., Khusainov, R. & Ndzi, D. (2007). Email classification and summarization: A machine learning approach. *IET Conference on Wireless, Mobile and Sensor Networks (CCWMSN07).*

Ayodele, T., Zhou, S. & Khusainov, R. (2009). Email grouping and summarization: An unsupervised learning technique. *WRI World Congress on Computer Science and Information Engineering.*

Chaipornkaew, P., Prexawanprasut, T. & McAleer, M. (2017). You've got email: A workflow management extraction system. *Journal of Reviews on Global Economics*, *6*, 342-349.

Chan, J., Koprinska, I. & Poon, J. (2004). Co-training with a single natural feature set applied to email classification. *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, 586-589.

Kiritchenko, S. & Matwin, S. (2001). Email classification with co-training. *Proceedings of the 2001 Conference of the Centre for Advanced Studies on Collaborative Research*, 1-10.

Kiritchenko, S. & Matwin, S. (2011). Email classification with co-training. *Proceedings of the 2011 Conference of the Centre for Advanced Studies on Collaborative Research*, 301-312.

Katakis, I., Tsoumakas, G. & Vlahavas I. (2006). *E-mail Mining: Emerging Techniques for E-Mail Management.* Web data management practices: Emerging techniques and technologies, Idea Group Publishing.

Kushmerick, N. & Lau, T. (2005). Automated email activity management: An unsupervised learning approach. *Proceedings of the 2005 International Conference on Intelligent User Interfaces*, 67-74.

Mihajlo, G., Halawi, G., Karnin, Z. & Maarek, Y. (2014). How many folders do you really need? Classifying email into a handful of categories. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 869-878.

Nenkova, A. & Bagga, A. (2003). Email classification for contact centres. *Proceedings of the 2003 ACM Symposium on Applied Computing*, 789-792.

Prexawanprasut, T. & Chaipornkaew, P. (2017). Email classification model for workflow management systems. *Walailak Journal of Science and Technology*, *14*(10), 783-790.

Schuff, D., Turetken, O., D'Arcy, J. & Croson, D. (2007). Managing e-mail overload: Solutions and future challenges. *Computer*, *40*(2), 31-36.

Taliby, R., Dean, R., Milner, B. & Smith, D. (2006). Email classification for automated service handling. *Proceedings of the 2006 ACM Symposium on Applied Computing*, 1073-1077.

Tam, T., Ferreira, A. & Lourenco, A. (2012). Automatic foldering of email messages: A combination approach. *Proceedings of the 34th European Conference on Advances in Information Retrieval*, 232-243.

Yelupula, K. & Ramaswamy, S. (2008). Social network analysis for email classification. *Proceedings of the 46th Annual Southeast Regional Conference on XX*, 469-474.

Yoo, S., Yang, Y. & Carbonell, J. (2011). Modelling personalized email prioritization: Classification-based and regression-based approaches. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, 729-738.

This article was originally published in a special issue, entitled: "**Management Information, Decision Sciences and Cognate Disciplines",** Edited by Chia-Lin Chang, Michael McAleer & Wing-Keung Wong