WALAILAK JOURNAL
http://wjst.wu.ac.th Information Technology

Email Classification Model for Workflow Management Systems*

Takorn PREXAWANPRASUT* and Piyanuch CHAIPORNKAEW

College of Innovative Technology and Engineering, Dhurakij Pundit University, Bangkok 10210, Thailand

(*Corresponding author's e-mail: takorn.pre@dpu.ac.th, piyanuch.chw@dpu.ac.th)

Received: 26 September 2016, Revised: 5 April 2017, Accepted: 30 May 2017

Abstract

The researchers observed and studied the business operations of 3 startup businesses in the export/import field. It was found that employees and their clients mostly communicate via email. Therefore, crucial business data are conveyed in email contents. Whenever employees need to find information, the first place they look for such data is email. The owners of businesses are concerned about this issue, so they proposed to buy a new workflow management system to help in managing their business transactions. The difficulty of implementing the new workflow management system is in migrating existing emails into the system. A new workflow management system should also be able to classify any incoming emails into categories. The researchers noticed that there were some keywords that frequently occurred in email contents in the same categories. Therefore, the researchers implemented a program to categorize the emails based on the words found in email messages. There are 2 parameters which affect the accuracy of the program. The first parameter is the number of words in a database compared to the sample emails. The second parameter is an acceptable percentage to classify emails. The results of this research demonstrated that the number of words in a database compared to the sample emails should be 9, and the acceptable percentage to categorize emails should be 30 %. When this rule was applied to categorize 8,751 emails, the accuracy of this experiment was approximately 73.6 %. The next phase is to order emails in each category based on their characteristics. Finally, the program extracts essential data from structured emails and prepares them for the new workflow management system.

Keywords: Business operations, startup business, import/export field, email, business data, workflow management system, business transactions, migrating

Introduction

Nowadays, the number of startup businesses has become much larger than before. The researchers participated as members of 3 startup businesses in the import/export field. It was found that they established their own businesses by separating from their former companies. There were 2 main reasons why they needed to separate and quit from their former companies. Firstly, they believed that they could win a greater market share in the import/export field because it was becoming larger and larger every day. Secondly, they thought that their former companies had ignored some groups of smaller clients. Since the former companies were quite big, they needed to serve their high-level clients first. Subsequently, some lower-level clients were ignored. Therefore, some of the lower-level clients had moved to another competitive company who served them better.

The researchers selected 3 startup businesses in the import/export field as the sample set. The owners of these startup businesses had quit their jobs from the same former company and launched their

*Presented at 1st International Conference on Information Technology: October 27th - 28th, 2016

own businesses. The researchers found that their business operations had 3 main characteristics: (1) They retained customer loyalty from their clients even though they had quit from their former company and had launched their own companies, (2) They needed to manage a large number of daily documents/emails; therefore, they needed applications to help them work faster and better, (3) They contacted their customers and employees mostly via emails. They also employed these emails, which were stored in the mail server, as a database. For example, when they wanted to find specific data, emails were the first place to look. At the first stage of starting their businesses, the number of emails was not large. However, when the scale of business has expanded, the number of emails also rose. The business owners then needed applications to manage their companies' activities. One of the most popular applications is the workflow management system.

The researchers selected 8,000 emails from 3 startup businesses. Only emails that were written in English were taken into consideration because sentences in English are easier to separate into words than emails in Thai. By reviewing some of these emails, the researchers noticed that some keywords specified the type of work, such as sales, transportation, billing, or shipping. The purpose of this research is to implement an application to classify the types of email to implement a future workflow management system

Literature review

There are many researches that mention the clustering and classification of email content. One of those is the work of Alsmadia and Alhamib [1]. The authors illustrated 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. They also stated that this algorithm best fit the bi-language text. In their paper, 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 folder. The general pre-defined folders could be mailbox, sent, trash, etc. Moreover, email servers or applications could allow users to add new folders for specific purposes based on their NGram algorithm.

Another research paper about email classification is the work of Katakis *et al*. [2]. They stated that Machine Learning and Data Mining could be used as tools to automate email managing tasks which could be much better than other conventional solutions. They also discussed the particularity of email content and what special treatment it required. In addition, there were some interesting email mining applications like mail categorization, summarization, automatic answering and spam filtering also presented in their paper. In their experiment, they created an application to classify email based on many techniques, such as the Naïve Bayes Classifier and Support Vector Machines. Ayodele *et al*. [3] presented the design and implementation of a system to group and summarize email messages. Their system considered the subject and content of email messages to classify emails based on user activities and produced summaries of each incoming message with an unsupervised learning approach. They claimed that their framework could solve the problem of email overload, congestion, difficulties in prioritizing and difficulties in finding previously archived messages in the mail server.

Another interesting research topic is email grouping and summarization. Ayodele *et al*. [4] presented the design and implementation of an application to categorize and summarize email content. Their system extracted the subject and content of email messages to classify the emails based on user activities to auto generate a summary of each incoming message. They stated that their framework could solve the main problems, such as email overload, difficulties in prioritizing, and email congestion. Their framework also performed successful processing of new incoming messages. Another interesting concept is automated email activity management as in the research of Kushmerick and Lau [5]. They developed email applications that provide high-level support for structured activities in e-commerce. They defined formal activities as finite-state automata, which correspond to the status of the process, and where transitions represent messages sent between participants. They proposed several unsupervised machine learning algorithms in this paper, and evaluated a collection of e-commerce emails. The work of Schuff *et al.* [6] also mentioned email classification. They implemented effective e-mail management tools which

treated 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 claimed that their application provided automatic filtering, clustering, and a new user interface. Their system employed a large number of emails as an effective knowledge management tool rather than a source of information overload.

Materials and methods

This research was designed in 2 phases as shown in **Figure 1**. The first phase was to select 512 emails randomly from the email server. Employees then had to manually classify all of these selected emails into 4 categories: (1) sales, (2) shipping, (3) billing, and (4) transportation. After categorizing the selected emails, the sentences in each email were separated into words and the frequency of each word was counted as shown in **Figure 2**. All results were stored in the database. These results could be defined as rules to classify the category of email. The next process was to verify these rules. The researchers prepared another 270 emails to test the defining rules. After these rules had been accepted, the researchers classified another 8,751 emails into 4 categories by using this program. The emails were generated from July 8, 2015 to July 31, 2016 as shown in **Figure 3**. When all emails were assigned into their categories, all data were extracted in the second phase of this email classification model.

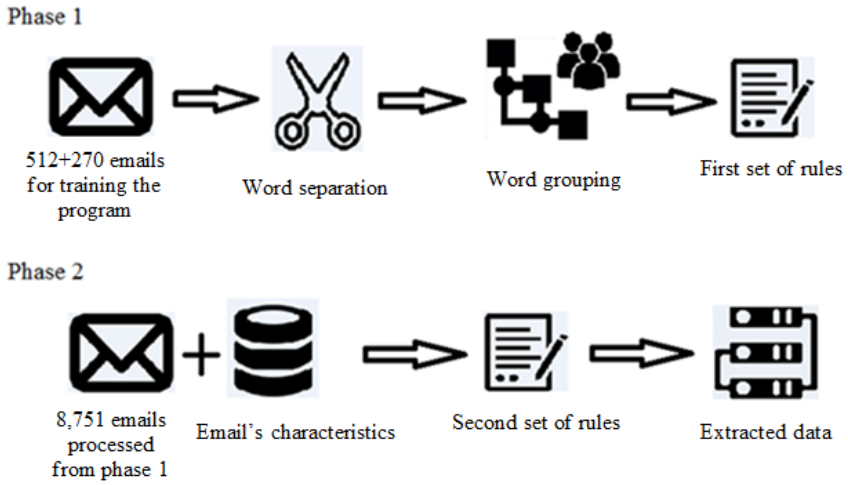
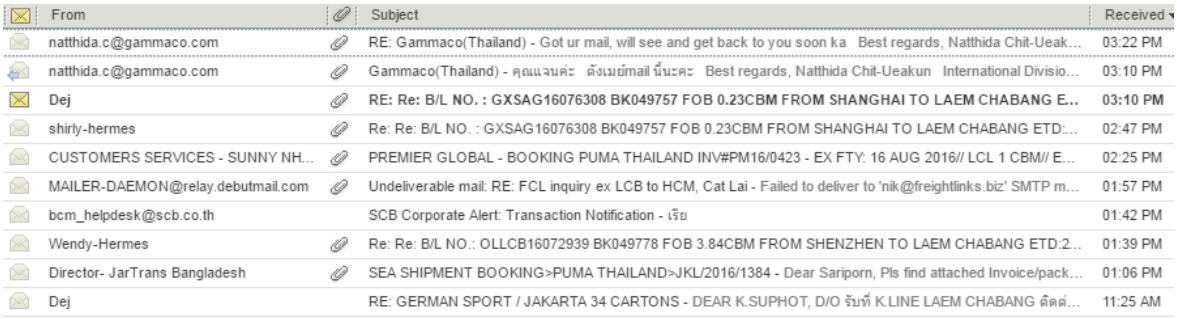


Figure 1 Two phases of the email classification model.

	about	Above	according	acknowledge	adding	advise	after	again	ahead
Sale	4	0	4	0	0	1	1	3	1
Shipping	1	2	0	0	5	0	3	1	0
Billing	2	0	1	1	1	1	0	2	1
Transportation	1	1	0	1	1	0	1	1	0

	all	already	and	any	anytime	apply	around	arrange	Arrival	as attache
Sale	3	0	8	2	1	2	0	5	3	0
Shipping	4	2	5	4	1	0	1	0	4	2
Billing	2	2	1	4	2	1	0	1	5	2
Transportation	2	1	4	0	0	0	2	4	5	0

Figure 2 Example of results from the word separation process.



From	Subject	Received
natthida.c@gammaco.com	RE: Gammaco(Thailand) - Got ur mail, will see and get back to you soon ka Best regards, Natthida Chit-Ueak...	03:22 PM
natthida.c@gammaco.com	Gammaco(Thailand) - คุณแจนค่ะ ดังเมย์mail นี้นะคะ Best regards, Natthida Chit-Ueakun International Divisio...	03:10 PM
Dej	**RE: Re: B/L NO. : GXSAG16076308 BK049757 FOB 0.23CBM FROM SHANGHAI TO LAEM CHABANG E...**	**03:10 PM**
shirly-hermes	Re: Re: B/L NO. : GXSAG16076308 BK049757 FOB 0.23CBM FROM SHANGHAI TO LAEM CHABANG ETD:...	02:47 PM
CUSTOMERS SERVICES - SUNNY NH...	PREMIER GLOBAL - BOOKING PUMA THAILAND INV#PM16/0423 - EX FTY: 16 AUG 2016// LCL 1 CBM// E...	02:25 PM
MAILER-DAEMON@relay.debutmail.com	Undeliverable mail: RE: FCL inquiry ex LCB to HCM, Cat Lai - Failed to deliver to 'nik@freightlinks.biz' SMTP m...	01:57 PM
bcm_helpdesk@scb.co.th	SCB Corporate Alert: Transaction Notification - เรีย	01:42 PM
Wendy-Hermes	Re: Re: B/L NO.: OLLCB16072939 BK049778 FOB 3.84CBM FROM SHENZHEN TO LAEM CHABANG ETD:2...	01:39 PM
Director- JarTrans Bangladesh	SEA SHIPMENT BOOKING>PUMA THAILAND>JKL/2016/1384 - Dear Sariporn, Pls find attached Invoice/pack...	01:06 PM
Dej	RE: GERMAN SPORT / JAKARTA 34 CARTONS - DEAR K.SUPHOT, D/O รับที่ K.LINE LAEM CHABANG ติดต่...	11:25 AM

Figure 3 Examples of emails in the mail server.

The second phase was to extract 8,751 emails which were already processed in phase 1. In this phase, the program would first reorder emails in each category. The emails and their characteristics were then gathered and stored in the database. Email characteristics included client data such as name, address, telephone number, and other data such as document number, contact person, and sales person. All these data were considered to determine the relationships between emails and the program extracted the specific data from these relationships. The final stage was applying the extracted data to implement a future workflow management system.

Data analysis

The first stage was to format all emails in a text file format and then import them into the program. The program separated the words in each sentence. As mentioned in the previous section, the selected emails were in English. The researchers decided to consider only emails in English because the separation of words in English was easier than in Thai. The algorithm to separate the words in English was based on the spaces between them. After separating the words in the email content, the program counted the frequency of each word in the emails. All words and their frequencies were stored in the database as shown in **Table 1**.

One example of emails determined to be in the shipping category stated “Dear Kae, Pls check subject shipment, kindly confirm B/L draft as soon as possible. Tks.” After this email was read by the program, 14 words were extracted. The program also counted the frequency of each word presented in the email message. In this case, the frequency of each word is 1 except the word “as”. There were 2 occurrences of “as” in the email content.

The researchers defined the mechanism to classify 8,751 emails into 4 categories: (1) Sales, (2) Shipping, (3) Billing, and (4) Transportation. This mechanism was based on the words found in emails compared to the words in the database for each category. There were 2 parameters in this experiment. The first parameter was the number of words in the selected database. For instance, the researchers needed to find out whether the first 5 words or the first 10 words in a database should be considered to obtain greater accuracy in email classification. The second parameter was the number of matching percentages which should be the most suitable to determine the category of email.

According to the data in **Table 2**, some emails could not be grouped because the percentage of matching words was less than the specified criteria. Additionally, the second criterion of this table was the top 10 words in a selected database. To obtain better results, the researchers needed to change these 2 criteria. As shown in **Table 3**, the top 20 words in a database were considered instead of the top 10 words. The researchers also set a matching criterion to be 20 %. As a result, some groups of output were different from **Table 2**. The first difference was the No. 4 group of emails. In **Table 2**, Email No. 4 could not be grouped but it was possible to categorize it in **Table 3**, the Transportation group. The second difference was the No. 6. group of emails. It was grouped as Transportation in **Table 2**, but in **Table 3** it could be in either Sales or Transportation.

Table 1 Top 10 words found in emails in 4 categories.

Sales		Shipping		Billing		Transportation	
Word	**Frequency**	**Word**	**Frequency**	**Word**	**Frequency**	**Word**	**Frequency**
agent	86	Shipment	154	consignee	109	loading	75
volume	54	Scheduled	87	shipper	76	destination	64
#NAME of CUS	51	ETA	76	#NAME of PORT	74	carrier	58
product	48	#Date format	48	revise	52	loader	55
shipment	32	BL	42	#NAME of CUS	45	#NAME of PORT	42
#NAME of CITY	30	HBL	40	scheduled	32	shipment	41
process	27	shipper	38	departed	18	#Date format	28
confirm	25	port	38	shipment	12	ETD	24
week	18	#NAME of CUS	35	#Date format	12	co-loader	22
#Date format	16	confirm	32	arrived	11	scheduled	22

Table 2 Grouping result based on the top 10 words and the 40 % matching criteria.

No. of Emails	Matching Percentage				Grouping Result
	Sales	**Shipping**	**Billing**	**Transportation**	
1	0.00 %	20.00 %	10.00 %	0.00 %	N/A
2	40.00 %	0.00 %	10.00 %	0.00 %	Sales
3	40.00 %	10.00 %	0.00 %	50.00 %	Transportation
4	0.00 %	0.00 %	20.00 %	10.00 %	N/A
5	0.00 %	0.00 %	60.00 %	0.00 %	Billing
6	10.00 %	10.00 %	10.00 %	40.00 %	Transportation

Table 3 Grouping results based on the top 20 words and the 20 % matching criteria.

No. of Emails	Matching Percentage				Grouping Result
	Sales	**Shipping**	**Billing**	**Transportation**	
1	5.00 %	10.00 %	15.00 %	0.00 %	N/A
2	25.00 %	10.00 %	5.00 %	0.00 %	Sales
3	20.00 %	10.00 %	5.00 %	30.00 %	Transportation
4	10.00 %	0.00 %	10.00 %	20.00 %	Transportation
5	5.00 %	0.00 %	35.00 %	0.00 %	Billing
6	20.00 %	10.00 %	5.00 %	20.00 %	Sales or Transportation

The experimental data from both **Tables 2** and **3** illustrate that there were 2 main factors that affected the grouping results. The first factor was the number of words in the selected database. The second factor was an acceptable matching percentage. Therefore, the researchers tested another 270 emails by changing the criteria for these 2 factors each time. The experimental results are plotted in **Figure 4**.

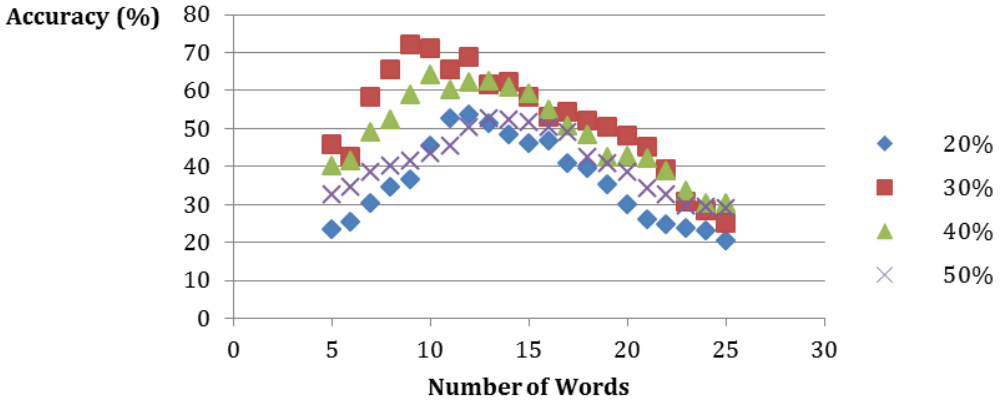


Figure 4 Accuracy (%) of email classification.

Results and discussion

The results shown in **Figure 4** demonstrate that the different accuracy levels would change when the number of words and matching percentage changed. The purpose of this experiment was to discover the suitable values for both the number of words and the matching percentage. The number of words in a database to be considered was adjusted from 5 to 25. The matching percentage was also adjusted to be 20, 30, 40 and 50 %, accordingly. According to the results from **Figure 4**, the highest accuracy level of email classification happened when the number of matching percentage was 30 % and the number of words in a database to be considered was 9. Therefore, the researchers defined these criteria in the program. Afterwards, the program was used to categorize the other 8,751 emails. These emails were classified into 4 groups: (1) Sales, (2) Shipping, (3) Billing, and (4) Transportation as shown in **Table 4**.

Table 4 The number of emails in each category.

Sales	Shipping	Billing	Transportation	Unclassified	Total
1,865	1,056	1,452	2,068	2,310	**8,751**

According to **Table 4**, the program could not categorize all emails. It defined only 6,441 emails from a total of 8,751 emails, which represents 73.6 % of the total emails. There were 2,310 emails which could not be grouped in this experiment. To improve the results of this experiment, it may be necessary to include other factors that are not included in this research. One example of possible factors could be the importance level of each word (the weight of each word) in a database. For example, the words which were most-frequently found in emails should be put at a higher importance level than the ones that were found less frequently.

When all emails were categorized into groups (Sales, Shipping, Billing, and Transportation), the next phase was to analyze the characteristics of these emails. The 4 main characteristics were: (1) the date

the email had been sent, (2) the document number shown in the email, (3) dates mentioned in the email message, and (4) the client's name mentioned in the email message. The program collected these characteristics and defined the order of events and relationships for all emails in each category. These results can be applied in future workflow management systems to improve the daily operations of businesses. The researchers intend to implement this workflow management system in future work. The program results are shown in **Figures 5** and **6**.

Figure 5 Example of program results after grouping, event ordering, and the inclusion of email characteristics.

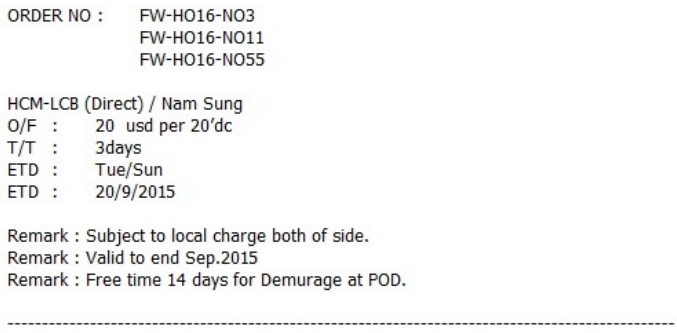


Figure 6 Example of the extracted data from phase 2.

Conclusions

From this experiment, there were 2 factors which impacted the accuracy of email classification. The first factor was the number of words in a selected database. The second factor was an acceptable matching percentage. After running the program with different numbers for these 2 factors, the results illustrated that the most suitable value for the number of words in a database was 9. In addition, the 30 % matching percentage provided the highest accuracy level. The results also demonstrated that the high

accuracy levels fall in the range of the number of words between 9 and 15 in every criterion of the matching percentage.

As mentioned earlier, this experiment selected all emails in English, so the researchers needed to exclude some words which are often found in every kind of email, such as 'and', 'or', 'thanks', 'dear', and 'please'. Since these kinds of word could not be used as criteria to define the category of email, the researchers needed to make decisions about the selection of words that should not be processed by the program. Moreover, there were some other words which should not be used as criteria in email classification. In our case, the examples of these words were FREIGHTLINKS, STARSHIP, and HERMESINT'L. These words were actual clients' names. Therefore, the researchers defined them as clients' names in the database, and the program would not apply them as criteria for email classification in the first phase. Even though the clients' names were excluded from the criteria of the first phase of email classification, they would be needed in the second phase of email classification to define the order of events and also the relationships of each email for future workflow management systems.

References

[1] I Alsmadia and I Alhamib. Clustering and classification of email contents. *J. King Saud Univ. Comput. Inform. Sci*. 2015; **27**, 46-57.

[2] I Katakis, G Tsoumakas and I Vlahavas. *Web Data Management Practices: Emerging Techniques and Technologies*. Idea Group Publishing, Pennsylvania, 2006, p. 220-43.

[3] T Ayodele, R Khusainov and D Ndzi. Email classification and summarization: A machine learning approach. *In*: Proceedings of the IET Conference on Wireless, Mobile and Sensor Networks. Shanghai, China. 2007, p. 805-8.

[4] T Ayodele, S Zhou and R Khusainov. Email grouping and summarization: An unsupervised learning technique. *In*: Proceedings of the WRI World Congress on Computer Science and Information Engineering. Los Angeles, USA, 2009, p. 575-9.

[5] N Kushmerick and T Lau. Automated email activity management: An unsupervised learning approach. *In*: Proceedings of the 2005 International Conference on Intelligent User Interfaces. San Diego, USA, 2005, p. 67-74.

[6] D Schuff, O Turetken, JD Arcy and D Croson. Managing e-mail overload: Solutions and future challenges. *Computer* 2007; **2**, 31-6.