You've Got Email: A Workflow Management Extraction System

Piyanuch Chaipornkaew[1], Takorn Prexawanprasut[1,*] and Michael McAleer[2-6]

[1]*College of Innovative Technology and Engineering, Dhurakij Pundit University, Thailand*

[2]*Department of Quantitative Finance, National Tsing Hua University, Taiwan*

[3]*Discipline of Business Analytics, University of Sydney Business School, Australia*

[4]*Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam, Netherlands*

[5]*Department of Quantitative Economics, Complutense University of Madrid, Spain*

[6]*Institute of Advanced Sciences, Yokohama National University, Japan*

Abstract: Email is one of the most powerful tools for communication. Many businesses use email as the main channel for communication, so it is possible that substantial data are included in email content. In order to help businesses grow faster, a workflow management system may be required. The data gathered from email content might be a robust source for a workflow management system. This research proposes an email extraction system to extract data from any incoming emails into suitable database fields. The database, which is created by the program, has been planned for the implementation of a workflow management system. The research is presented in three phases: (1) define suitable criteria to extract data; (2) implement a program to extract data, and store them in a database; and (3) implement a program for validating data in a database. Four criteria are applied for an email extraction system. The first criterion is to select contact information at the end of the email content; the second criterion is to select specified keywords, such as tel, email, and mobile; the third criterion is to select unique names, which start with a capital letter, such as the names of people, places, and corporates; the fourth criterion is to select special texts, such as Co. Ltd, .com, and www. The empirical results suggest that when all four criteria are considered, the accuracy of a program and percentage of blank fields are at an acceptable level compared with the results from other criteria. When four criteria are applied to extract 7,340 emails in English, the accuracy of this experiment is approximately 68.66%, while the percentage of blank fields in a database is approximately 68.05. The database created by the experiment can be applied in a workflow management system.

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

"I do love email … I'm really good at email."

Elon Musk

1. INTRODUCTION

Given an increase in business competition in recent years, email has become an indispensable business tool to drive organization processes. Essential business information can be distributed to many people by one click of an email button.

Emails contain valuable information that can be used to improve business operations. As employees in the organizations always communicate with their clients via emails, substantial customer data are likely to be included in email content. In many organizations, a large number of historical emails are used to perform data mining in order to extract valuable knowledge, which is hidden inside the emails.

The authors were invited to be part of three startup import/export business organizations in Thailand. These three companies have the same outstanding issue, which is a method to deal with a large number of emails. As shown in Figure **1**, the usage of data storage for one account is approximately 17 GB. The business owners realize that they need computer systems to help them use the data included in email content.

As the authors explored content in emails, much data could be used as an important resource for future business plans, such as customer names, customer telephones, and company names. In order to access this useful information, two specific programs are required. One program is for extracting data from email content and storing them in a database, while the other program is for validating the results in a database.

Some email content might not be necessary to take into consideration, so suitable criteria should be

*Address of correspondence to this author at the College of Innovative Technology and Engineering, Dhurakij Pundit University, Thailand; Tel: +66 2954 7300; Fax: +66 2954 8651; E-mail: takorn.pre@dpu.ac.th

JEL: J24, O31, O32, O33.

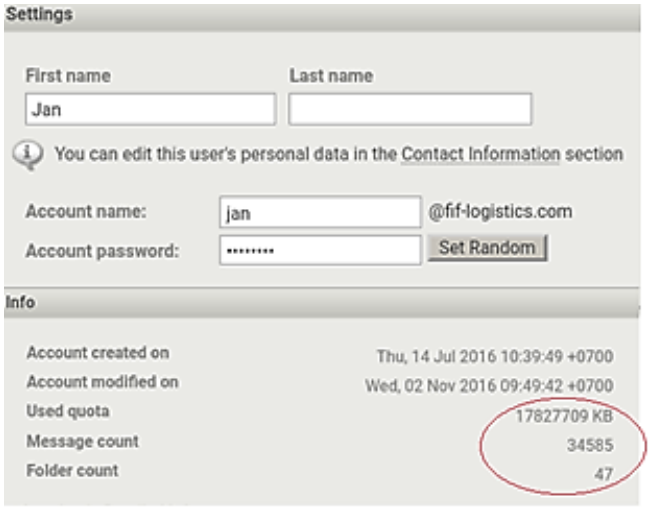


Figure 1: Usage of Data Storage for One Account.

defined for data extraction. Based on daily business emails, the patterns of email content are roughly consistent. As all emails are created while employees operate their businesses, the responsibility to define criteria for email extraction should be assigned to employees. In order to verify the accuracy of the email extraction program, employees are also responsible for validating the results, which are in database fields.

The purpose of this paper is to extract data from email content based on four criteria, which are defined by employees. The extracted data will then be stored in suitable database fields, which are applied in a workflow management system. In order to verify the accuracy of the extraction program, the specific program is also implemented for assigned employees to validate the data in database fields.

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

2. LITERATURE REVIEW

Email summaries are mentioned in many research papers. One of the interesting topics is summarizing email conversations with clue words (see Carenini *et al*. (2007)). Researchers have suggested a method called CWS to summarize conversations in emails. The framework applies two techniques namely, using: (i) a fragment quotation graph to capture an email conversation; and (ii) clue words to measure the importance of sentences in conversation summaries. Researchers have claimed that their method provides better summaries of email conversations than existing methods.

Muresan *et al*. (2001) and Tzoukermann *et al*. (2001) have applied the same approach to summarize emails, namely a combination of linguistic and machine learning techniques. The paper shows that linguistic techniques and machine learning can extract high quality noun phrases for purposes of providing a summary of email messages.

Hailpern *et al*. (2014) also address the email summary issue. In order to summarize the content of email attachments, a novel email attachment summary system was created, namely AttachMate. The system can perform summaries, and automatically insert the summary into the text of the email.

Summarizing text conversations is also proposed by Carenini and Murray (2001). The research presents various natural language processing (NLP) techniques for mining and summarizing text conversations. Nomoto and Matsumoto (2012) also present a novel approach that exploits the diversity of concepts in text. A diversity-based approach is a principled generalization of the Maximal Marginal Relevance criterion (MMR), which selects a sentence in such a way that it is both relevant to the query and has the least similarity to sentences selected previously.

In addition to a diversity-based approach in Nomoto and Matsumoto (2012), the researchers also apply an information-centric approach where the quality of summaries is judged not in terms of how well they match human-created summaries but in terms of how well they represent their source documents in text categorization.

Another approach concerned with text categorization is Bekkerman *et al*. (2003). The paper presents an approach that combines distributional clustering of words and a Support Vector Machine (SVM) classifier. A Support Vector Machine is based on the concept of decision planes that define decision boundaries. The technique performs classification by finding the hyperplane that maximizes the margin between the two classes. The paper suggests that a combination of these two methods provides higher performance in text categorization.

A clustering of words is presented in Chrupała (2012), who proposes an unsupervised approach to POS tagging, which is the process of marking up a

word in a text as corresponding to a particular part of speech, based on both its definition and its context. The approach is a hierarchical clustering of the word types and is defined as an agglomerative clustering algorithm, which is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.

Another type of word clustering is given in Baker and McCallum (1998), who describes the application of distributional clustering for document classification. The approach clusters words into groups based on the distribution of class labels associated with each word. Feature space refers to the n-dimensions where variables live, and is used often in machine learning literature because a task in machine learning is feature extraction. This method can compress the feature space much more aggressively, while maintaining high document classification accuracy.

Shunyao *et al*. (2010) consider text clustering with important words using normalization. The paper proposes a novel method to extract important words from the subject and keywords of articles. A normalization method is then proposed to scale the dataset so that more accurate results can be achieved. In Hui *et al*. (2003), a rule-based context-dependent word clustering method is introduced. The paper defines rules based on various domain databases and word text orthographic properties. The experiments show that such rule-based word clustering improves the accuracy of extracting bibliographic fields from references.

3. MATERIALS AND METHODS

The research below is conducted in three phases, as shown in Figure **2**. The first phase selects 7,340 emails in English from the email server. As the email server backup function cannot be managed remotely, company employees are responsible to provide data from their email inbox. The emails were generated from Aug 8, 2016 to Jan 31, 2017. Employees then analyze the email content and define criteria to extract data. There are four criteria to select data from email content, namely (1) contact information at the end of the email; (2) keywords; (3) unique names; and (4) special text.

The second phase is implementing a program to gather specified data from email content based on the four criteria processed in phase 1. Then the program separates words, which are collected and then stored in the suitable database fields. This empirical database is designed for implementing a workflow management system. An example of the words gathered from email content and separated is given in Figure **3**. An example of contact information at the end of the email content is

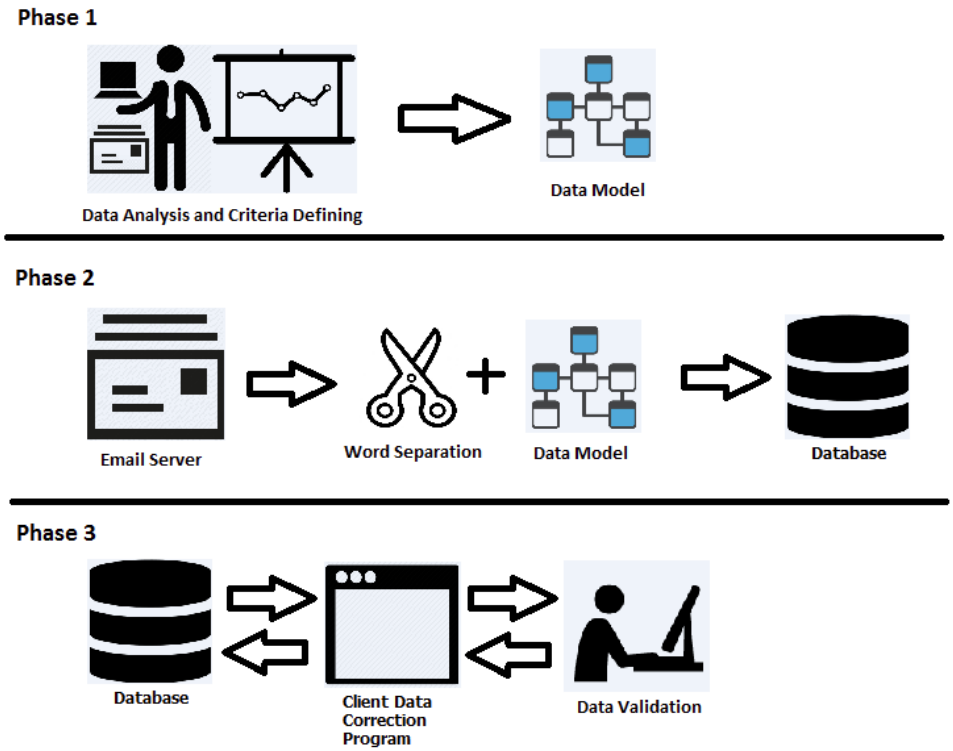


Figure 2: Three Phases of an Email Extraction System.

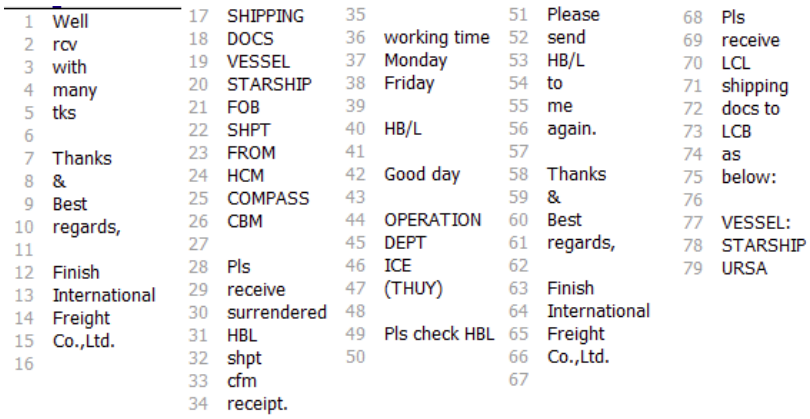


Figure 3: Example of Words Gathered from Email Content.

in Figure **4**, and an example of data in a database is presented in Figure **5**.

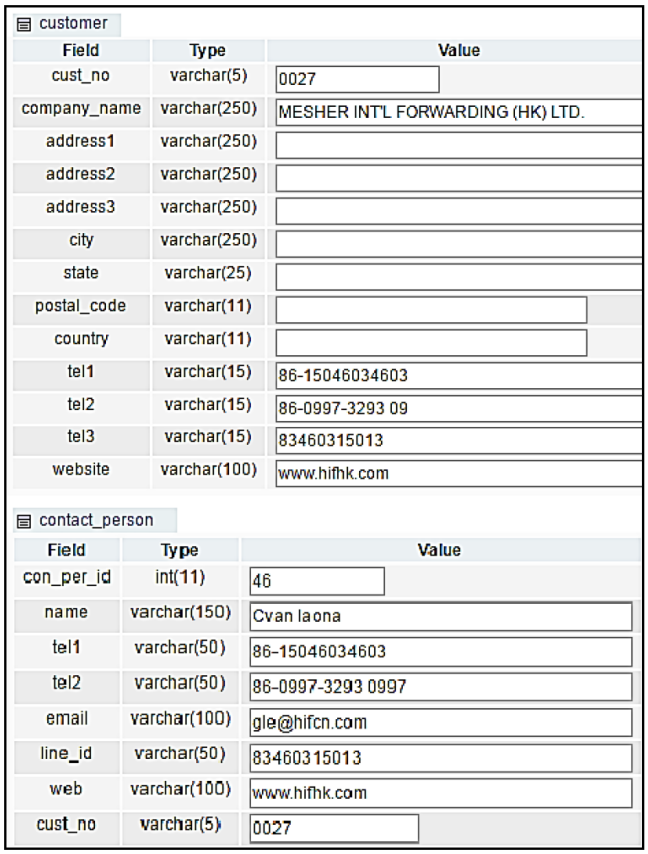
Figure 4: Example of Contact Information at the End of an Email.

The last phase is validating the data in a database. In order to validate the data, another program is created. The program provides specified employees to retrieve, correct, and restore data in a database. The employees need to fill out the missing data, correct the incorrect data, and change the data that are not in the corresponding database fields. An example of missing data in a database is shown in Figure **6**. The data, which are completely validated, are marked as accurate data. Every correction is recorded in a special table, known as history, in a database. An example of the program for employees to verify the data is shown in Figure **7**. The history table in a database is presented in Figure **8**.

4. DATA ANALYSIS

In order to evaluate the efficiency of the email extraction program, there are two proposed factors. The first factor is the number of blank fields in the database, which cannot be filled in by the program. The blank fields, indicated by <to be added>, are shown in Figure **7**. The second factor is the data that are filled incorrectly by the program. The data edited by employees are shown in Figure **8**. The data in a database are extracted based on four criteria, as mentioned in phase 1, namely (1) contact information at the end of the email; (2) keywords; (3) unique names; and (4) special text in the email content.

As the accuracy of the program depends on four criteria, the authors tested the program by changing the criteria. The empirical results are shown in Table **1**, and are plotted in Figure **9**. The lowest percentage of blank fields happens when the four criteria are considered. However, the percentage accuracy when all criteria are



customer

Field	Type	Value
cust_no	varchar(5)	0027
company_name	varchar(250)	MESHER INT'L FORWARDING (HK) LTD.
address1	varchar(250)	
address2	varchar(250)	
address3	varchar(250)	
city	varchar(250)	
state	varchar(25)	
postal_code	varchar(11)	
country	varchar(11)	
tel1	varchar(15)	86-15046034603
tel2	varchar(15)	86-0997-3293 09
tel3	varchar(15)	83460315013
website	varchar(100)	www.hifhk.com

contact_person

Field	Type	Value
con_per_id	int(11)	46
name	varchar(150)	Cvan Iaona
tel1	varchar(50)	86-15046034603
tel2	varchar(50)	86-0997-3293 0997
email	varchar(100)	gle@hifcn.com
line_id	varchar(50)	83460315013
web	varchar(100)	www.hifhk.com
cust_no	varchar(5)	0027

Figure 5: Example of a Database.

cust_no	company_name	address1	address2	address3	city	state	postal_code	country	tel1	tel2
0023	ShuShen Electronics	906-907, 09th Floor,Section B Taipingyang Commer...			Shenzhen			China		

Figure 6: Example of Missing Fields, Address2, Address3, State, Postal_code, Tel1, Tel2.

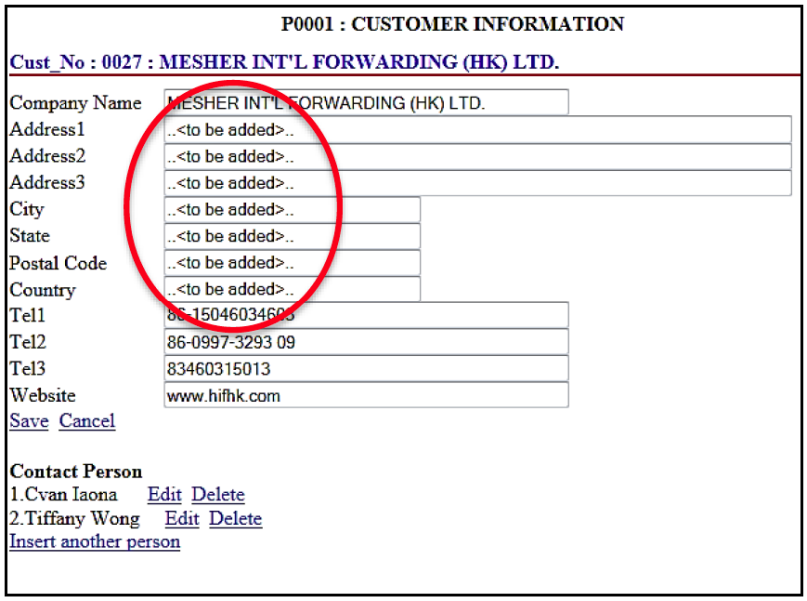


Figure 7: Example of the Program for Employees to Verify Data.

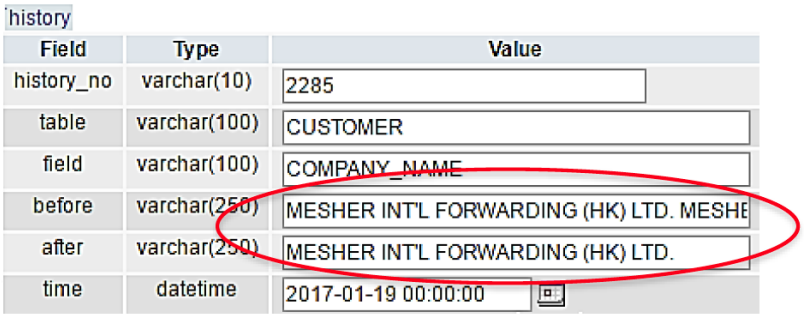


Figure 8: History Table in a Database.

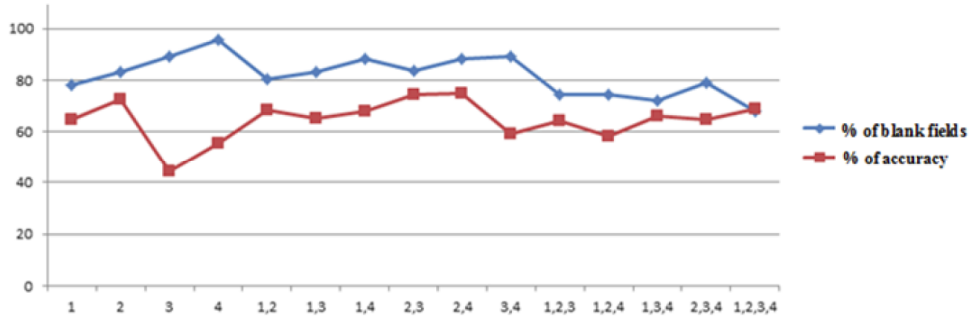


Figure 9: Results of Criteria Combinations.

considered is less than the others. The lowest percentage of blank fields is 68.05, while the percentage accuracy is 68.66, as shown in Table **1**. The highest percentage accuracy occurs when criteria

Table 1: Results of the Program with Alternative Criteria

Case Number	Number of criteria	% of blank fields	% accuracy
1	1	78.23	64.50
2	2	83.28	72.45
3	3	89.12	44.22
4	4	95.87	55.54
5	1, 2	80.55	68.33
6	1, 3	83.21	65.21
7	1, 4	88.12	68.11
8	2, 3	83.56	74.52
9	2, 4	88.17	74.65
10	3, 4	89.12	59.22
11	1, 2, 3	74.44	64.14
12	1, 2, 4	74.21	58.45
13	1, 3, 4	72.12	65.88
14	2, 3, 4	78.98	64.68
15	1, 2, 3, 4	68.05	68.66

Note: Number of criteria defined as: (1) contact information at the end of email; (2) keywords; (3) unique names; (4) special text. Results are calculated from 7,340 emails in English.

numbers 1 and 2 are considered. The highest percentage accuracy is 74.65, as shown in Table **1**. As criterion number 2 yields suitable results in terms of accuracy and percentage of blank fields, criteria number 2 should be considered as having greater accuracy in email extraction.

Another proposed strategy to support this idea is to evaluate the results for different combinations of criteria. The results of combining the criteria are shown in Table **2**, criterion number 2 has the highest percentage accuracy at 68.24, and the lowest percentage of blank fields at 78.91.

5. RESULTS AND DISCUSSION

The results in Tables **1** and **2** demonstrate that changing the number of criteria can affect the accuracy of the extraction program. The purpose of this section is to determine suitable criteria to be considered for email extraction. According to Table **1** and Figure **9**, the number of criteria is adjusted in fifteen cases, where each case is composed of different groups of criteria. Table **1** demonstrates that the highest accuracy level of email extraction occurs when criteria numbers 2 and 4 are selected. The highest percentage accuracy is approximately 74.65, while the percentage of blank fields is approximately 88.17. The lowest percentage of blank fields occurs when all four criteria are considered. The lowest percentage of blank fields is approximately 68.05, while the percentage accuracy is 68.66. Both sets of results indicate that criterion number 2 has the greatest impact in terms of the accuracy of the program.

In order to support the idea that criterion number 2 should be considered more favorably than the others, the paper provides further analysis, which are

Table 2: Results of Alternative Criteria

Criteria Combinations	% of blank fields	% accuracy
1: 1,2 – 1,3 – 1,4 – 1,2,3 – 1,2,4 – 1,3,4 – 1,2,3,4	77.37	65.41
2: 1,2 – 2,3 – 2,4 – 1,2,3 – 1,2,4 – 2,3,4 – 1,2,3,4	78.91	68.24
3: 1,3 – 2,3 – 3,4 – 1,2,3 – 1,3,4 – 2,3,4 – 1,2,3,4	79.83	63.32
4: 1,4 – 2,4 – 3,4 – 1,2,4 – 1,3,4 – 2,3,4 – 1,2,3,4	81.83	64.40

Note: Number of criteria defined as: (1) contact information at the end of the email; (2) keywords; (3) unique names; and (4) special text. Results are calculated from 7,340 emails in English.

combinations based on each criteria, as shown in Table **2**. The highest percentage accuracy is 68.24, while criterion number 2 is considered. However, the percentage of blank fields when criterion number 2 is considered is greater than the other cases. Consequently, it is not possible to conclude that criterion number 2 is the most suitable for email extraction.

According to the interviews of employees who are responsible for verifying the data, they prefer an email extraction program to store the data in a database rather than leave the database fields blank. The employees mentioned that, although the data that are stored are sometimes incorrect, employees still use such data for other database fields. Therefore, the lowest percentage of blank fields would seem to be the most suitable result for the employees.

6. CONCLUSION

There are four criteria to select data from email content, namely (1) contact information at the end of the email; (2) keywords; (3) unique names; and (4) special text. The paper examined whether the number of criteria has an impact on the accuracy of email extraction. After running the program with different groups of criteria, the results indicated that the highest accuracy percentage is 74.65 when criteria numbers 2 and 4 were considered. The results also demonstrated that the lowest percentage of blank fields, at 68.05, occurred when all four criteria were selected. The results from criteria combinations showed that criterion number 2 provided the highest percentage accuracy, at 68.24.

Although criteria numbers 2 and 4 should be selected together to gain the highest percentage accuracy, this case provided a greater number of blank fields. According to the considered views of company employees, the lowest percentage of blank fields was preferred. In order to follow their suggestions, all four criteria should be considered to yield the lowest number of blank fields.

All emails were selected from three startup businesses, but the results are not presented separately for each company. In order to improve the results, future research might examine the extracted data for each company based on their own email content, as each company might have different patterns of email content, thereby leading to different outcomes based on different databases.

As this paper has focused on email content, future research could apply this approach for other types of data, such as product details, sales, or employees. The empirical database in the paper is designed for a workflow management extraction system to improve the daily operations of businesses. Future research will implement the workflow management extraction system in practical applications, especially in business, finance and marketing.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to the Executive Vice-President of Finish International Freight Co. Ltd, as well as another two anonymous companies which cannot be mentioned because of confidentiality. The companies provided very useful information and insights to conduct this research. Thanks also to Khun Natthicha Phonjan and Khun Sariporn Plipon, who assisted greatly to edit and verify the accuracy of the program. 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. The authors are also grateful for the helpful comments and suggestions of Chia-Lin Chang.

REFERENCES

Baker, D. and McCallum, A. (1998), Distributional clustering of words for text classification, *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, August 1998, pp. 96-103. https://doi.org/10.1145/290941.290970

Bekkerman, R., El-Yaniv, R., Tishby, N., and Winter, Y. (2003), Distributional word clusters vs. words for text categorization, *Journal of Machine Learning Research Archive*, Volume 3, March 2003, 1183-1208.

Carenini, G. and Murray, G. (2012), Methods for mining and summarizing text conversations, *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, August 2012, pp. 1178-1179. https://doi.org/10.1145/2348283.2348529

Carenini, G., Raymond, T., and Xiaodong, Z. (2007), Summarizing email conversations with clue words, *Proceedings of the 16th International Conference on World Wide Web*, May 2007, pp. 91-100. https://doi.org/10.1145/1242572.1242586

Chrupała, G. (2012), Hierarchical clustering of word class distributions, *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, June 2012, pp. 100-104.

Hailpern, J., Asur, S., and Rector, K. (2014), AttachMate: Highlight extraction from email attachments, *Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology*, October 2014, pp. 107-116. https://doi.org/10.1145/2642918.2647419

Hui, H., Manavoglu, E., Giles, C., and Hongyuan, Z. (2003), Rule-based word clustering for text classification, *Proceedings of the 26th Annual International ACM SIGIR Conference on*

Research and Development in Information Retrieval, August 2003, pp. 445-446.

Muresan, S., Tzoukermann, E., and Klavans, J. (2001), Combining linguistic and machine learning techniques for email summarization, *Proceedings of the 2001 Workshop on Computational Natural Language Learning*, Volume 7, July 2001, pp. 1-8. https://doi.org/10.3115/1117822.1117837

Nomoto, T. and Matsumoto, Y. (2001), A new approach to unsupervised text summarization, *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, September 2001, pp. 26-34. https://doi.org/10.1145/383952.383956

Shunyao, W., Jinlong, W., Huy, V., and Gang, L. (2010), Text clustering with important words using normalization, *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, June 2010, pp. 393-394.

Tzoukermann, E., Muresan, S., and Klavans, J. (2001), GIST-IT: Summarizing email using linguistic knowledge and machine learning, *Proceedings of the Workshop on Human Language Technology and Knowledge Management*, July 2001, pp. 1-8. https://doi.org/10.3115/1118220.1118231

Received on 16-02-2017 Accepted on 13-05-2017 Published on 09-06-2017

DOI: https://doi.org/10.6000/1929-7092.2017.06.35