Mining Business Process Information from Email Logs for Business Process Models Discovery

Diana Jlailaty
Université Paris-Dauphine,
PSL Research University,
CNRS, [UMR 7243], LAMSADE,
75016 Paris, France
diana.al-jlailaty@dauphine.fr

Daniela Grigori
Université Paris-Dauphine,
PSL Research University,
CNRS, [UMR 7243], LAMSADE,
75016 Paris, France
daniela.grigori@dauphine.fr

Khalid Belhajjame
Université Paris-Dauphine,
PSL Research University,
CNRS, [UMR 7243], LAMSADE,
75016 Paris, France
khalid.belhajjame@dauphine.fr

***Abstract*—Exchanged information in emails' texts is usually concerned by complex events or business processes in which the entities exchanging emails are collaborating to achieve the processes' final goals. Thus, the flow of information in the sent and received emails constitutes an essential part of such processes i.e. the tasks or the business activities. An email can be harvested for understanding the undocumented business process information it contains. Our goal in this work is to recast emails into a resource of business-oriented information. We describe a framework that is constituted of several analytical approaches able to extract such kind of information from email logs i.e. transforming an email log into an event log. The efficiency of all approaches is studied by applying different experiments on the open Enron email dataset.**

***Index Terms*—Email Analysis, Business Process Models, Text Mining, Process Instances, Business Activities**

I. Introduction

Email is by and large the first and the most popular professional communication and social medium [1]. It is a reliable, confidential, fast, free and easily accessible form of communication. Exchanging emails becomes essential when applying tasks in organizational processes necessitates the involvement of multiple individuals. Assigning tasks, asking for more information, reporting results - all these activities are enacted via email messages. Therefore, such email messages necessarily contain process-related information that refer to the business process under execution.

However, email analysis from a Business Process Management (BPM) perspective has not been thoroughly studied in the literature. Some of the existing works allow the identification of email activities among a predefined set of activities [5], [4], [3]. The email analyzer developed by Van der Aalst [6] necessitates the user interference to extract a process instance from an email log. Hence, until recently and up to our knowledge, none of the previous works has tackled the problem of extracting business process information from emails *automatically* without any a priori knowledge for the goal of business process models discovery.

In this work, we aim to analyze the unstructured data in emails to harvest the undocumented business process information from email logs i.e. event logs. In an event log, events are characterized by some attributes. Each event corresponds to an activity (associated to an activity label) that is executed in the process (associated to a Process Identifier), where multiple events (ordered by their timestamps) can be linked together as a process instance (associated to a Process Instance Identifier). Transforming email logs into event logs allows us to produce business process models using the available process mining tools. The produced business process models can provide a clear overview on the processes and the activities in a user email log which facilitates the organization and retrieval of emails. In this work, we develop a framework that includes different approaches contributing in the following:

- An approach that can automatically find, for each email, the business process topic it belongs to i.e. extraction of the Process Identifier (ProcessID) for each email.
- A process instance discovery approach that can automatically find the business process instance an email belongs to i.e. extraction of the Process Instance Identifier (ProcessInstanceID) for each email.
- An approach that automatically extracts multiple business activities from emails and that annotates the elicited activities i.e. extraction of the activity labels in an email.
- A preliminary approach that can estimate the real occurrence time of an event or email activity i.e. extraction of the activity occurrence timestamp.
- The efficiency of all the above approaches is evaluated using multiple email folders from Enron email dataset [2].

In this paper, we first start by providing a brief study on the related works in section II. An overview on the overall framework is presented in section III. Sections IV, IV, VI, and VII explain the phases of our framework. Finally, the work is concluded in section VIII.

II. Related Work

The common objective of the related works presented in this section is to categorize emails into a set of classes (folders, topics, importance, main activities). In the work of Alsmadi et al. [1], a large set of emails is used for the purpose of folder classifications. Five classes are proposed to label the

[1] http://onlinegroups.net/blog/2014/03/06/use-email-for-collaboration/

[2] https://www.cs.cmu.edu/ enron/

nature of emails: Personal, Job, Profession, Friendship, and Others. Another work by Yoo et al. [7] develop a personalized email prioritization method using a supervised classification framework. The goal is to model personal priorities over email messages, and to predict importance levels for new messages using standard Support Vector Machines (SVMs) as classifiers. In the work of Bekkerman et al. [2], they represent emails as bag-of words (vectors of word counts) to classify them into a predefined set of classes (folders). In the work of Faulring et al. [5], they classify tasks contained within sentences of emails from 8 predefined set of classes of tasks.

In our work, we overcome the limitation of specifying a predefined set of process topic and activity classes. Our approach is able to cope with the diversity of business topic and activity types that can exist in emails. In contrast to some previous works, we automatically discover and label all topic and activity types presented in an email. In addition, instead of dealing with a single activity type per email, we work on discovering multiple business-oriented activities in an email.

III. Framework Overview

In this section, we present the overall framework developed in this paper. It is composed of three main components. Figure 1 shows an overview of the components of the framework. The framework takes as an input the email log. The first component is for **Process Topic Discovery** where each email is associated to a business process topic. The second component is for **Process Instances Discovery** where each email is associated to a business process instance. The third component if for **Process Activities Discovery** where each email is associated to a set of process activities.

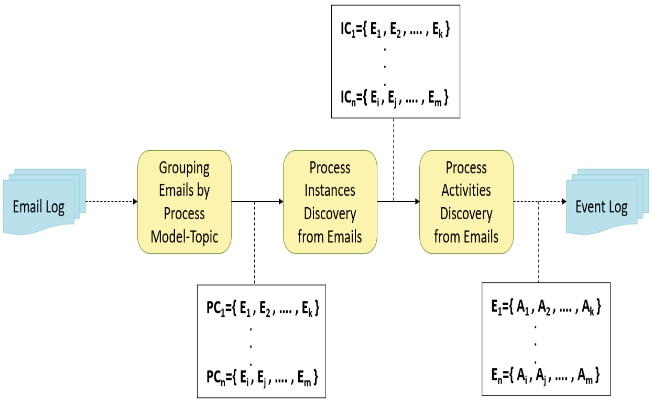
IC_1={ E_1, E_2, …., E_k }
IC_n={ E_i, E_j, …., E_m }
Email Log
Grouping Emails by Process Model-Topic
Process Instances Discovery from Emails
Process Activities Discovery from Emails
Event Log
PC_1={ E_1, E_2, …., E_k }
PC_n={ E_i, E_j, …., E_m }
E_1={ A_1, A_2, …., A_k }
E_n={ A_i, A_j, …., A_m }

Fig. 1. The overall framework.

IV. Process Topic Discovery

A. Email Log Preprocessing

Each email in an email is represented by some attributes describing it: email subject, sender, receiver, email body, and email timestamp. Knowing that email text falls under the category of unstructured data, one must take considerable time to preprocess this data with fixed fields so that they can be queried, quantified, and analyzed with data mining techniques. Four main steps are applied in this component:

1) Data Cleansing: removing stopwords, whitespaces, and puctuations, or stemming.
2) Data Representation: representing emails bodies and subjects as numerical Term Frequency-Inverse Document Frequency matrices that can be used in the analysis.
3) Verb-Nouns Extraction: we consider that the verb-noun pairs are likely to be candidates of being business activities.

B. Clustering Emails According to their Process Topics

In this section, our objective is to group the emails into clusters according to what process model they are concerned by. If we fetch the emails of a researcher, we can detect that his/her emails are concerned by different processes such as scheduling a meeting or organizing a conference, etc..

Since in our case, we do not have any apriori knowledge about the business process topics available in the email log, we use an unsupervised machine learning method which is the clustering. In particular, hierarchical clustering is used to group emails in clusters based on their similarity. Agglomeratively (using the complete linkage), the clusters are fused together according to the chosen similarity measurement technique. We try two different methods for semantic similarity measurement between the selected features of emails (1) Latent Semantic Analysis which is used to map words occurring in emails into concepts. and (2) Word2vec which is used to calculate the similarities between emails according to the context of their words. The hierarchy is cut such that emails belonging to same process model are clustered together. The output of this cut is a set of clusters $\{PC_1, PC_2, PC_3, ..., PC_n\}$, where each cluster PC_i contains a set of emails related to the same process model topic. PC_i and subsequently the emails contained in it are associated to a ProcessID.

V. Process Instances Discovery

In process mining, there exist the main terms: *business process model* and the *business process instance*. A process instance is a specific occurrence or execution of a business process model. Each email log can contain processes of different topics. Let us suppose that in an email log, the "meeting scheduling" process topic exists. An employee may exchange emails with two different entities for scheduling different meetings. Thus, multiple email exchanges take place for this purpose. These email exchanges represent the different executions or occurrences. In this section, we work on analyzing email texts for identifying for each email the process instance it belongs to. This is mainly done by choosing attributes and features from email texts that help in distinguishing between emails of different process instances and in grouping of the same process instances. We work on choosing the best distance function that consists of a combination of attributes for clustering emails into process instances.

For discovering business process instances from email logs, we start from the previously obtained process topic clusters $\{PC_1,\ PC_2,\ PC_3,\ ...,\ PC_n\}$, where each cluster PC_i represents a business process topic. We aim to deduce for each business process topic cluster, the set of process instances it contains.

A. *Defining an Appropriate Distance Function*

In order to separate emails belonging to the same process model into different process instances, we apply a sub-clustering step on the already obtained process topics clusters. We illustrate the steps of this phase and the distance function definition by using an example which is concerned by applications for missions funding.

Example: Suppose that one of the obtained clusters contains emails about all applications of Ph.D students for "missions funding" process topic. Emails of the same process instance are supposed to revolve around some common names. Take as an example the emails exchanged between a student and the secretary for applying for a funding to attend the BDCSIntell 2019 conference in Versailles. Most of these emails bodies and subjects will include the named entities "BDCSIntell" or "Versailles" or "Paris". We claim that these named entities can be helpful in discovering which emails are related to the same process execution (same mission application). However, in some cases named entities will not be sufficient to distinguish instances. Suppose two emails are about applying to a travel grant for the same conference BDCSIntell but in two different years 2018 and 2019. These two emails are supposed to belong to different process instances. Using only the named entities, these two emails will be considered belonging to the same process instance. Thus, we decide to add another attribute which gives an indication about the time of sending an email. Although the named entities and the email timestamp have provided a good indication for separating emails into different instances, some cases have proven that these two attributes are not always sufficient. Suppose two different students are applying to the same conference BDCSIntell in the same year 2019. The named entities and the timestamps of the emails of these students will be similar, however, these emails belong to different process instances (for two different students). Therefore, we add a new attribute which is sender/receiver of an email which can separate emails as in the described case. We define the distance function as follows: we first define the similarity function and then derive the distance function.

$$Distance(E_{i_j}, E_{i_k}) = 1 - (w_1 \times Sim(NE_{i_j}, NE_{i_k}) + w_2 \times Sim(T_{i_j}, T_{i_k}) + w_3 \times Sim(SR_{i_j}, SR_{i_k})) \tag{1}$$

where E_{i_j} and E_{i_k} are two different emails j and k in the same process model cluster C_i. (NE_{i_j}, T_{i_j}, SR_{i_j}) and (NE_{i_k}, T_{i_k}, SR_{i_k}) are the named entities of the subjects and bodies, timestamps and sender/receiver of emails E_{i_j} and E_{i_k} respectively. Weights w_1, w_2, w_3 ($w_1 + w_2 + w_3 = 1$) represent the relative importance of named entities, timestamps and sender/receiver of emails, respectively.

B. *Clustering Emails Into Process Instances*

Using the above distance function, we calculate distances between all pairs of emails. Accordingly, hierarchical clustering is applied where we get the emails distributed on a hierarchical structure. We tried several cuts on the obtained hierarchy. We choose the one which provides the best clustering quality (according to clustering quality measures mentioned in the experimentation section). Each of the obtained clusters IC_i contains emails belonging to the same process instance. Every cluster is provided a process instance identifier.

VI. Process Activities Discovery

A *Business Process Model* is composed of a set of *Business Activities* enacted in a specific sequence to achieve a business goal. Each email is sent for the aim of requesting, canceling, confirming a specific task or set of tasks. One of the main attributes in the event log is the activity label. Therefore, in this section, we work on extracting activity labels from email logs. There are two main hypotheses for the extraction of business process activities from email logs. The first hypothesis is that each email contains one and only activity which is not always true. Therefore, we propose the second hypothesis in which we assume that an email can contain 0, 1 or more activities.

We build an approach that takes as input an email log and produces the set of business activity types it contains.

A. *Relevant Sentences Extraction*

The goal of this phase is to extract from each email the sentences that contain business activities or information about activities. To identify relevant sentences in an email, we use a classification technique that associates each email sentence with one of the following labels *Relevant* or *Non-Relevant*. We characterize each email sentence by a set of features that describe it: (1) Sentence position, (2) Sentence length, (3) Number of named entities, (4) Cohesion with centroid sentence, (4) Dissimilarity with greeting phrases, (5) Length of the sentence, (6) Similarity of the verb-nouns with process-oriented activities extracted from a repository of process models of different domains.

We build the training data by labelling the the sentences feature vectors. An expert decides whether the sentence is meaningful from a business-oriented perspective. The training dataset is used to train the classification model. Different classification techniques are used to obtain the best classification results.

B. *Activity Types Discovery*

The business activities elicited in the previous steps can be further processed to organize them by their *activity types*. Hierarchical clustering is applied to sentences containing process oriented verb-noun pairs (i.e. activities). The similarity between two sentences is calculated using the cosine similarity between the Word2vec vectors of their verb-noun pairs (i.e. activities). This phase will give as a result a set of clusters where each cluster contains sentences from different emails but with the same activity type ($\{AC_i\}$). For each cluster, we choose the top N verb-noun pairs mentioned in the activity cluster (for example N can be equal to 3). Then one of these verb-noun candidates can be chosen by an expert as a label for the cluster.

VII. Temporal Features Extraction

Each email is exchanged in a specified timestamp. As a first hypothesis, one can consider that the timestamp of an email activity is the same as that of the email it belongs to which is not always true. An email may contain business activities that were already applied, in progress activities, or activities that will be applied in the future. Therefore, we extract the temporal relation:

1) Between the email activities and the email timestamp: the objective is to temporally locate an email activity according to the email timestamp in which it occurs. To be accurate, we can divide the relation between the activity and the email timestamp into different categories. Possible categories are *Before*: in which the activity occurs before the email sending time, *Overlap* in which the activity occurs at the time the email is sent and *After* in which the activity will occur after the email sending time.
2) Between the email activities themselves: the objective here is to extract the intra temporal relations between the email activities using the email temporal expressions.

VIII. Conclusions and Perspectives

Throughout the sections of this paper, we described the main approaches that constitute the framework presented in this work. The components of the framework are mainly concerned by: (1) business process topic discovery for emails, (2) business process instances discovery for emails, (3) business activities discovery, and (4) preliminary estimation of the activity occurrence timestamp.

There exist several potential perspectives based on the obtained results such as building a recommendation system that can recommend activities based on received emails, or allowing the incremental learning for our system.

References

[1] Izzat Alsmadi and Ikdam Alhami. Clustering and classification of email contents. *Journal of King Saud University-Computer and Information Sciences*, 27(1):46–57, 2015.

[2] Ron Bekkerman. Automatic categorization of email into folders: Benchmark experiments on enron and sri corpora. 2004.

[3] Vitor R Carvalho and William W Cohen. Improving email speech acts analysis via n-gram selection. In *Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech*, pages 35–41. Association for Computational Linguistics, 2006.

[4] William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. Learning to classify email into speech acts. In *In Proceedings of Empirical Methods in Natural Language Processing*, 2004.

[5] Andrew Faulring, Brad Myers, Ken Mohnkern, Bradley Schmerl, Aaron Steinfeld, John Zimmerman, Asim Smailagic, Jeffery Hansen, and Daniel Siewiorek. Agent-assisted task management that reduces email overload. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 61–70. ACM, 2010.

[6] Wil MP van der Aalst and Andriy Nikolov. Emailanalyzer: an e-mail mining plug-in for the prom framework. *BPM Center Report BPM-07-16, BPMCenter. org*, 2007.

[7] Shinjae Yoo, Yiming Yang, Frank Lin, and Il-Chul Moon. Mining social networks for personalized email prioritization. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 967–976. ACM, 2009.