Mining Business Process Models from Email Logs

Diana Jlailaty[1] · Daniela Grigori[1] · Khalid Belhajjame[1]

[1]Paris Dauphine University, Department of Computer Science, LAMSADE, Paris, France

Problem

- Undocumented business process information contained in email logs.
- The notion of process is absent in email management systems.

Motivation

- How many times a specific process (or activity) is applied?
- Which groups do similar work?
- What is the average duration taken by a process?
- Which process involves specific entities?

Contributions

- Extracting and analyzing business process information from an email log.
- Transforming the email log into an event log.
- Deducing business process models from email logs.

Proposed Approach

We follow a 4-phased approach which takes as an input an email log and produces a set of business process models.

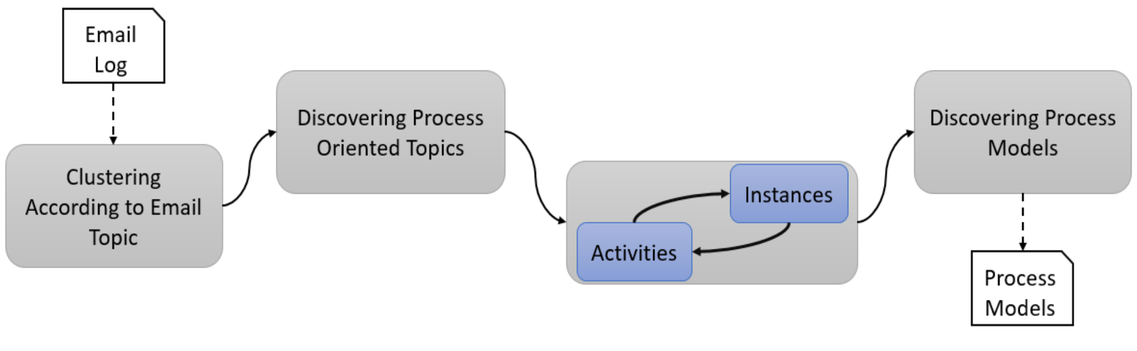


Figure 1. The overall approach

Phase 1: Topic Clustering

In this phase, emails are clustered according to which topic they belong to. Emails belonging to the same topic are grouped together.

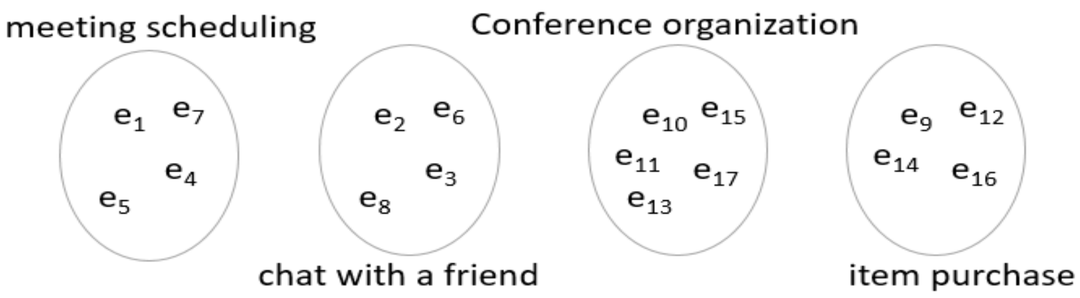


Figure 2. Obtained clusters

However, we are only interested in the process-oriented topics(emails) such as "meeting scheduling" or "conference organization".

Phase 2: Discovering Process-Oriented Topics

We differentiate between process and non process-oriented topics using some defined heuristics

- Process-oriented verb-noun pairs are specified in process model repositories.
- An ontology is built out of these pairs.
- If the emails contained in a specific cluster include such process-oriented verb-noun pairs, then the cluster topic is process-related.
- Another heuristics can help such as the existence of IDs in the email.

We can deduce the process ID for each email. Each process-oriented topic will be considered as a separate process characterized by a model (to be discovered).

Phase 3: Discovering Process Instances and Activities

- Process Activity Discovery:

(a) Sub-clustering is applied. Each sub-cluster corresponds to an activity.

(b) Activity labeling: For each sub-cluster, activity labels are provided by examining the most occurring n-grams.

- Process Instance Discovery:

(a) A distance function is defined which considers: named entities, timestamp, and sender/receiver. Each attribute is provided a different weight according to its importance in calculating distances accurately.

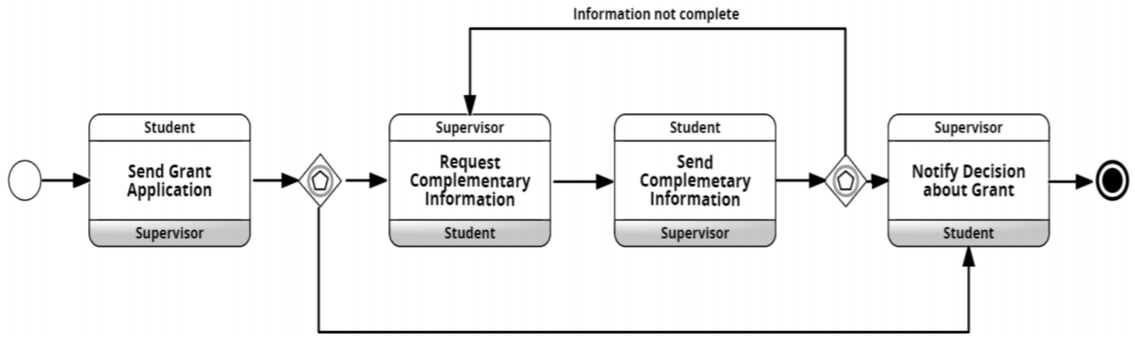
(b) We define the following distance function:

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

For the moment, process instance and activity discovery are done separately. We will study the relational problem between the two steps. We aim to iterate through both steps to ensure the best process instance and activity discovery.

Study 4: Process Model Discovery

We obtained for each email, the process ID, process instance ID and the activity label. The obtained data can be an input for the process mining tools to obtain the business process models hidden in the email log.



Future Work

We are working on the relational problem between process activity and process instance discovery phases. We will also test our approach on Enron dataset containing 500,000 emails.

Contact

Email: diana.jlailaty@gmail.com, Telephone: +33 6 09 21 66 62