Facilitating Business Process Discovery using Email Analysis

Matin Mavaddat
University of the West of England
Computer Science and Creative Technologies
Bristol, UK
Matin.Mavaddat@live.uwe.ac.uk

Stewart Green
University of the West of England
Computer Science and Creative Technologies
Bristol, UK
Stewart.Green@uwe.ac.uk

Ian Beeson
University of the West of England
Computer Science and Creative Technologies
Bristol, UK
Ian.Beeson@uwe.ac.uk

Jin Sa
University of the West of England
Computer Science and Creative Technologies
Bristol, UK
Jin.Sa@uwe.ac.uk

***Abstract*— Extracting business process models from stakeholders in large organizations is a very difficult, if not impossible, task. Many obstacles such as tacit knowledge, inaccurate descriptions of processes and miscommunication prevent process engineers from ascertaining what the business processes actually are. Data sources that represent the communications can be a good candidate for facilitating the identification of the business processes. The proposed approach in this research is to find business process related emails, identify email message threads, and finally, tag them using conversation for action theory. The outcome of this method will be process fragment enactment models that can help process engineers both to validate their findings about the business processes, and also to understand better the vague and unclear parts of the processes.**

Keywords-Business Processes; Email Analysis; Process Mining; Semantic Similarity.

I. INTRODUCTION AND BACKGROUND

"The reality is often very different from what is modelled or what people think. The world is not a Petri Net." Van der Aalst [1]. The researcher's experience confirms this, while working on two requirement extraction and business process analysis projects for Saint John Ambulance and Intellect Publishing Ltd. during the past three years. He has always spent an extensive amount of time with the clients and tried to use conventional tools and techniques, as well as best practices, for business process modelling and analysis, but there are always many obstacles in the way in order to find the actual business processes that are being followed in an organization. These are almost the same as challenges that we face in knowledge acquisition: for example, limited memory, information processing biases, representativeness, communication problems and different perceptions [8]. In this research, we are trying to find out if by using the data in the email corpus of an organization in conjunction with the conversation for action and speech act theory we can create fragments of business process enactments to help process engineers understand organizational processes better. Due to privacy issues, one of the challenges in analysing an email corpus is to have access to one. For this research the researcher has access to Intellect Publishing's email corpus. Another challenge is to what extent the business processes are being carried out using email messages. The models that are created using the proposed approach are called process instance "fragment" models and based on the amount of "fragmentation" these models can be very useful or of little use. This fragmentation is directly related to the number of business processes that are being carried out using the email system either fully or partially. Intellect Publishing, like many other organisations [3], uses emails as its main means of communication. Therefore, a good portion of its journal production processes is carried out using email messages.

The rest of the paper is organized as follows. In Section II, the main differences of the proposed approach and the existing ones will be discussed. In Section III, the solution approach will be introduced, and finally, in Section IV, we will put forward our conclusions and future work.

II. STATE OF THE ART

To date only a small amount of work has been done in this field, such as the works of Van der Aalst et al. [2] and Cohen et al. [16]. In all of these proposed works and methods some assumptions have been made that make using these techniques almost impractical. Assumptions like: Spam free mailboxes (Here spam means all the emails that are unrelated to the business processes) and finding the email threads only from email meta-data or by analysing email subjects or by manually created tags. In the proposed solution in this research, none of these assumptions have been made. Email messages are automatically filtered using text categorisation techniques and analysing emails' content semantics is added to the subject and meta-data analysis in order to overcome the shortcomings of previously introduced

 ISBN: 978-1-61208-160-1

methods. In addition to the aforementioned points, none of the previous methods have used the conversation for action and speech act theory to justify the soundness of this idea that email analysis might be a good method for facilitating the business process identification.

III. SOLUTION APPROACH

The final goal of this research is to create a method both to extract business process related emails from an email corpus, and to create a model from them that shows the interaction between different role instances involved in the business process. This method may help the process engineers validate or better understand the processes that are elicited by other means. The method has three main stages: email categorisation, conversation network finding, and conversation network tagging.

A. Email categorization

The first process, which feeds directly from the email corpus, is the email categorization process. In email categorization, text classification techniques are used. Text classification is assigning automatically a text document to a predefined class [8]. Obviously each email corpus contains several different types of emails such as business-rule related emails and personal emails which might not contain any business process related information. So it is necessary that the email corpus be divided (classified) into two different classes: business process related and non-business process related.

Many different methods and techniques have been introduced to solve this type of problem such as Naïve Bayes [7] and Support Vector Machines [5]. The main challenge in solving this problem is to find the features that help to classify textual documents (emails) into business process related and non-business process related.

In order to use automatic categorization techniques and also in order to choose the best categorization algorithm, a number of email messages should be classified manually to create a training set and a test set. A text-mining tool named WEKA [13] can be used to test different text mining algorithms. The training set can be fed to the WEKA tool, which automatically trains itself using different text mining training algorithms, including the ones that were mentioned before. WEKA automatically extracts classifying features that distinguish business process related emails from non-business process related emails using different algorithms, then, using the test set and feeding it to the tool, the algorithm that best matches the human classification of the emails can be identified.

B. Conversation network finder

After classifying the emails and finding the business process related emails, the next step involves finding email threads inside the business process related emails. Email threads, or email trees, are related email conversations that have occurred about a similar topic, ordered by time. Email threads that are created from the classified set of emails can be a representation of a network of "conversation for action" introduced by Winograd [12]. Winograd put forth the conversation for action theory based on Searl's [10] speech act theory. He argues that business processes are networks of conversations that are happening inside and outside the organization about the organizational goals. They introduce the conversation for action diagram and believe that almost every network of conversation for action happens according to the pattern introduced in that diagram. He believes "speech acts are not individual unrelated events, but participate in large conversational structure." [13]

For example, one conversation for action network might start with a request, and a request is understood by the participant as having certain conditions of satisfaction, which shows how the hearer might react to the speaker's request. Finding these networks of conversation for action" should help us realize organization's business processes.

 ISBN: 978-1-61208-160-1

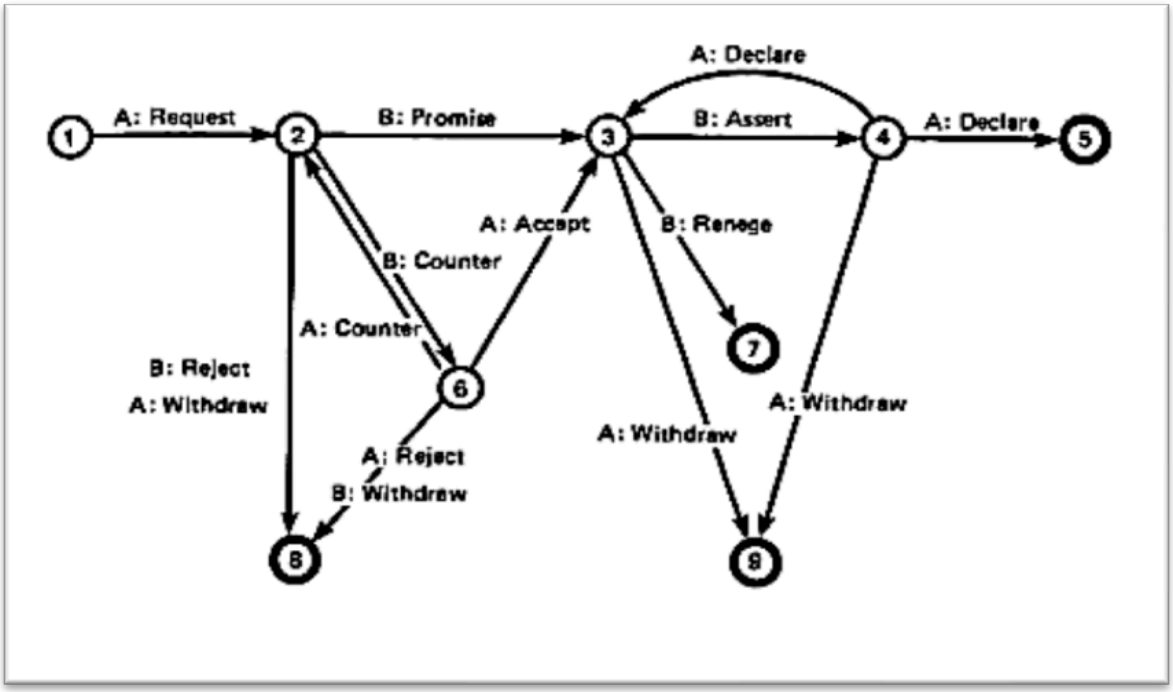

Figure 1. Conversation for Action diagram

One approach to find these networks of conversation is to use semantic similarity and some other heuristics [15]. By applying this process, the business process related set of emails turns into different threads, each representing one conversation network that is initiated by a role instance and has been continued until a mutual agreement.

One problem here is that if all the conversions have not happened via email, then it cannot be realized. This is the problem of incompleteness in process mining. This means we may not be able to model the whole process. This is why it is called a process instance (enactment) "fragment" extraction.

A Java application has been developed for this purpose and a function within this application is created to measure the semantic similarity of two email messages in order to realize if they are related. This is used in email thread finding. It is assumed that if an email is sent by a recipient of an original email after the date of the original email, and it is semantically similar to the original email, then it is part of the thread that is initiated by the original email.

A refined version of Vector Space Model algorithm [7] is used in this function. It means that each email is translated into a vector implemented in sparse matrix and the semantic similarity is measured using the multiplication of vectors.

The following refinements have been made to the original vector space model: first, the vectors are not two-dimensional but n-dimensional. The synonyms are added to the vector as the higher dimensions using WordNet [4]. It is possible in future to just add the context-related synonyms (by interpreting and using the WordNet as a lexical ontology and comparing it with the existing ontologies in the organisation); and second, Checking the words' spellings before adding them to the vector using spell checking algorithms.

The output of this function will be something like the following table for the simplified scenario below:

1. Matin sends an email to Stewart to request for a meeting and specifies his availability.
2. Stewart selects a date and time and sends an email to Jin to ask if she can make it too.
3. Jin sends an email to Stewart and accepts the date and time.
4. Stewart sends an email to Ian to see if he can make it at that date and time.
5. Ian responds that he can make it.
6. Stewart sends an email to Matin telling him the date and time of the meeting.

TABLE I. CONVESATION NETWORK FINDER OUTPUT TABLE

#	Sender	Recipient	Timestamp	Email ID
1	matin @uwe.ac.uk	Stewart@uwe.ac.uk	15-Jun-2010 14:00:00	132
2	Stewart @uwe.ac.uk	Jin@uwe.ac.uk	15-Jun-2010	139

 ISBN: 978-1-61208-160-1

			14:40:23	
3	Jin @uwe.ac.uk	Stewart@uwe.ac.uk	15-Jun-2010 15:39:30	150
4	Stewart @uwe.ac.uk	Ian@uwe.ac.uk	16-Jun-2010 09:30:30	180
5	Ian @uwe.ac.uk	Stewart@uwe.ac.uk	16-Jun-2010 12:30:12	200
6	Stewart @uwe.ac.uk	Matin@uwe.ac.uk	16-Jun-201015:30	250

The results created by this application were quite satisfactory when they were analyzed manually. The extracted threads were quite reasonable and close to the actual threads that were created manually.

C. *Conversation network tagging*

Up to this stage, the relation between role instances have been extracted; this shows the interaction between different roles instances to achieve a goal (mutual satisfaction), but the interactions are not labelled. This labelling can be done using the speech act theory [11] as this theory tries to define what the speaker intends to do by using words.

Searle [10] has set up the following classification of illocutionary speech acts: assertives, directives, commissives, expressives, and declarations.

By finding the illocutionary acts and the propositional content related to each email (or email paragraph if needed), we can label each interaction. For example, the result of this step is expressed by the following RAD-like diagram [9]:

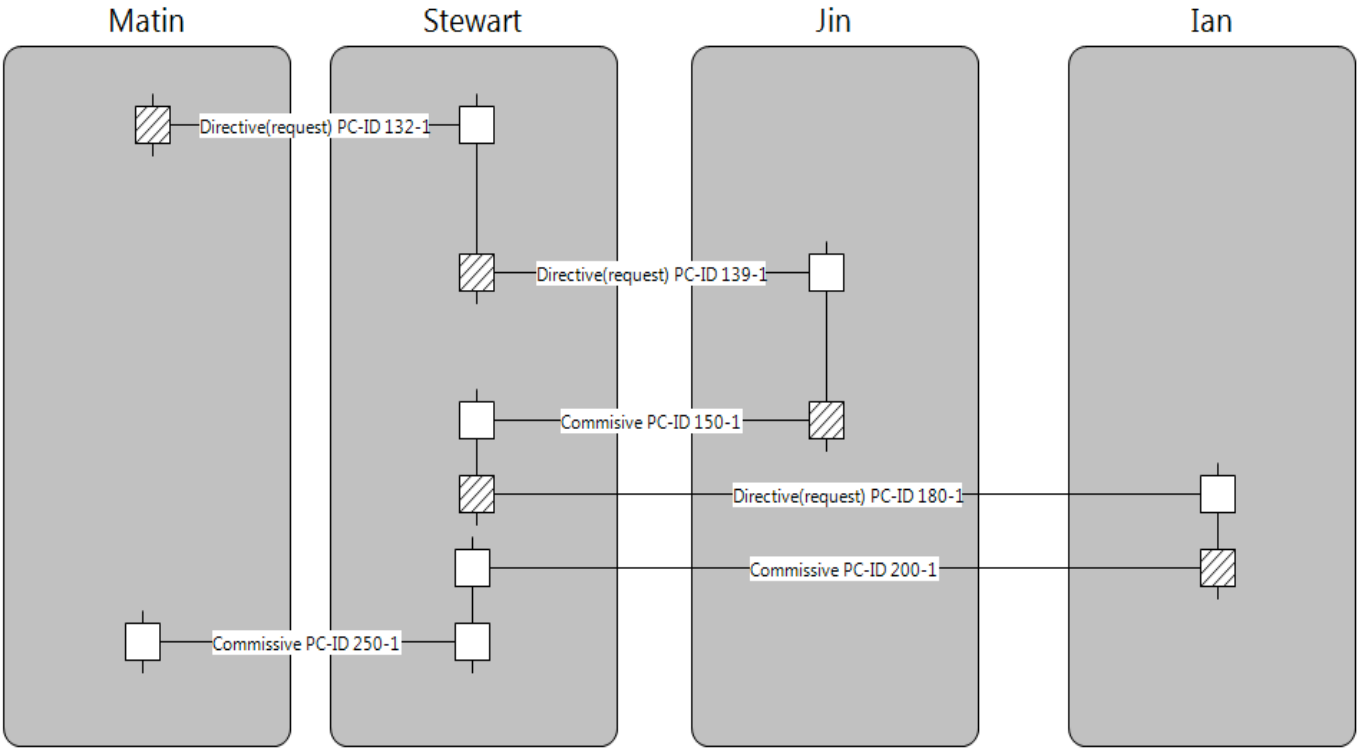

Figure 2. Conversation networks.

Fig. 2 shows the interactions between different roles instances and the interactions are tagged by the extracted illocutionary act of the relevant paragraph or email and also a link to the actual paragraph or email for manual reference and further analysis.

This model is an interaction fragment instance. By analysing these models and finding the similar patterns, we should be able to create business process fragment instance models. For instance, by analysing the interaction fragment instance models of different meeting scheduling occurrences and finding the patterns, we should be able to create meeting scheduling business process fragment models. These models may help business engineers to validate their understanding of the business processes, and also might clarify some vague fragments of the manually identified processes. Interpreting these models should not be difficult for the process engineers. RAD diagram notations are being used, the interactions are tagged by illocutionary acts that are quite

 ISBN: 978-1-61208-160-1

self-explanatory and each interaction is directly linked to the actual paragraph or email for further analysis if something is not clear enough.

IV. CONCLUSION AND FUTURE WORK

This paper proposed an approach to facilitate the identification of business processes of an organization Fig. 3. For the first sub process a database is created from the emails metadata and content, for analysis with WEKA for email categorization and for the second sub-process an application has been developed that identifies the threads using a modified version of the Vector Space Model algorithm. The third sub-process needs more research involving both tagging identified conversations using the "speech act theory", and using the "conversation for action" idea to find similar patterns and to create the fragments of business process instances from the interaction fragment instances.

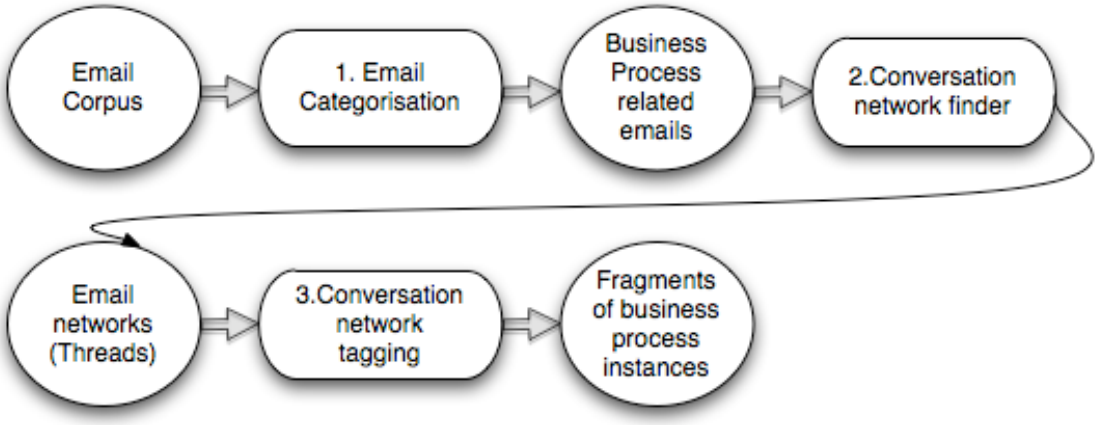

Figure 3. Solution Approach

REFERENCES

[1] Aalst, W. M. P., 2003. Challenges in business process analysis. Bulletin of the EATCS, 80, pp. 174-198.

[2] Aalst, W. M. P. and Nikolov, A., 2008. Mining e-mail messages: uncovering interaction patterns and processes using e-mail logs. International Journal of Intelligent Information Technologies, 4 (3), pp. 27-45.

[3] Ellis, C. A., 2000. An evaluation framework for collaborative systems, University of Colorado at Boulder, USA.

[4] Fellbaum, C. 1998. WordNet: An electronic lexical database (Language, Speech and Communication), MIT Press.

[5] Joachims, T. , 1998. Lecture Notes in Computer Science-Text categorization with Support Vector Machines: Learning with many relevant features, Springer, pp. 137-142.

[6] Liu, B., Lee, W. S., Yu, P. S., and Li, X., 2002. Partially supervised classification of text documents. ed. 19th International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., pp. 387 – 394.

[7] Manning, C. D., Raghavan P., and Schutz H., 2008. An introduction to Information Retrieval. Cambridge University Press.

[8] Mcgraw, K. L. and Harbison-Briggs, K., 1989. Knowledge acquisition principles and guidelines, Prentice- Hall.

[9] Ould, M., 2005. Business Process Management: A Rigorous Approach, BCS.

[10] Searle, J. R., 1975. A Taxonomy of Illocutionary Acts, in: Günderson, K. (ed.), Language, Mind, and Knowledge, Minneapolis, vol. 7, pp. 1-29.

[11] Searle , J. R., 1969. Speach Acts: An Essay in the Philosophy of Language Cambridge University Press.

[12] Winograd T. and Flores, F. , 1986. Understanding computers and cognition, Addison-Wesley.

[13] Winograd, T. , 1987. A language/action perspective on the design of cooperative work. Human-Computer Interaction, 3 (1), pp. 3-30.

[14] Witten, H. I. and Frank, E. , 2005. Data mining: practical machine learning tools and techniques. Morgan Kaufmann publishing.

[15] Yeh, J. and Harnly, A, "Email thread reassembly using similarity matching." Proc. Third Conference on Email and Anti-Spam (CEAS), 2006.

[16] Cohen, W., Carvalho, V., and Mitchell, T., 2004. Learning to classify email into "Speech Acts.". Association for Computational Linguistics,4(11), pp. 309-316, doi: 10.1002/asi.20427.

 ISBN: 978-1-61208-160-1