A Universal Significant Reference Model Set for Process Mining Evaluation Framework

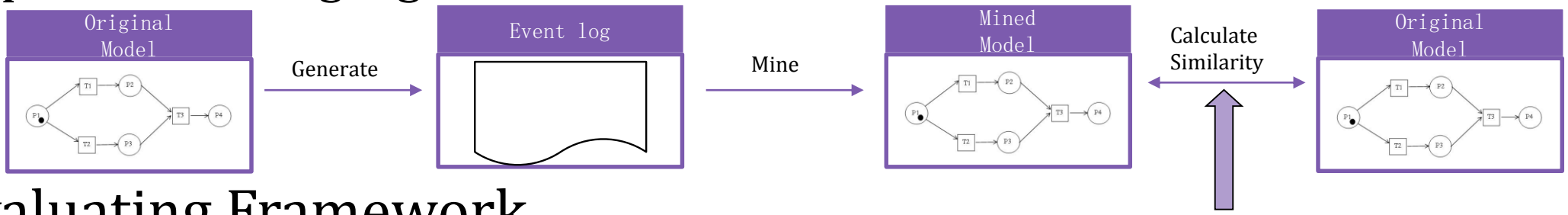
Qinlong Guo, Lijie Wen, Jianmin Wang,
Zizhe Ding, and Cheng Lv
Tsinghua University, China Mobile Communication Corporation

Content

Content

Background

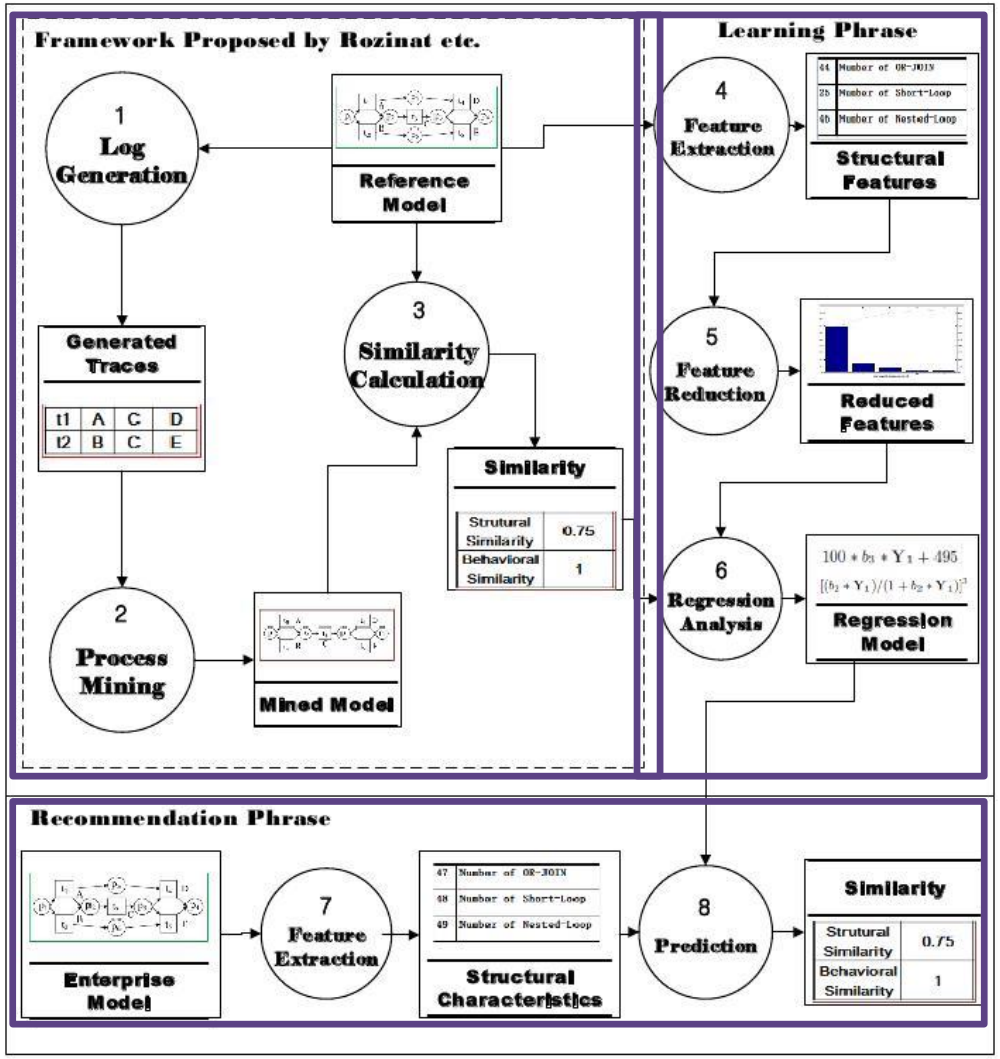
- Model Rediscoverability
  - Given a process model (we call it original model) and its corresponding event log, the *model rediscoverability* is to measure how similar between the original model and the process model mined by the process mining algorithm.



- Evaluating Framework
  - The effect of the recommendation is highly dependent on the quality of the reference models.
  - Nevertheless, choosing the significant reference models from a given model set is also time-consuming and ineffective.

Content

Evaluation Framework



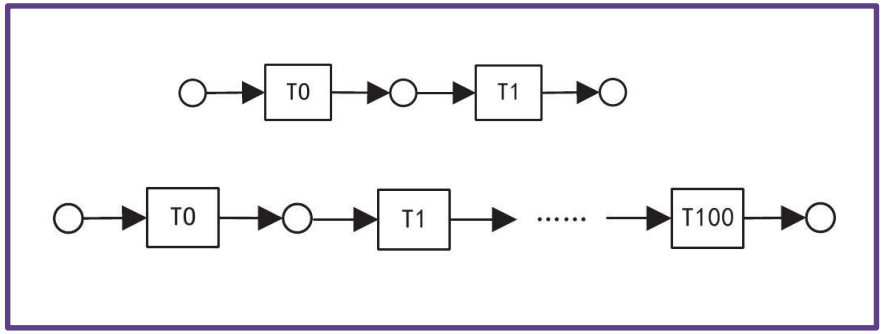
- **Model Rediscoverability**
- **Learning Phrase**
  - The reference models are used as training set to obtain the regression mode
- **Recommending Phrase**
  - The most suitable (i.e. best performing) mining algorithm can be recommended without performing the actual empirical benchmarking.
- **Improvement**
  - The reference models are the universal reference models rather than the process models selected from each dataset. This change makes the module 1-6 unrequired for each dataset.
  - The features extracted in the Feature Extraction module (the 4 and 7 module) are the selected 6 features rather than the 48 features.
- This Framework has been implemented in BeehiveZ 3.5*.

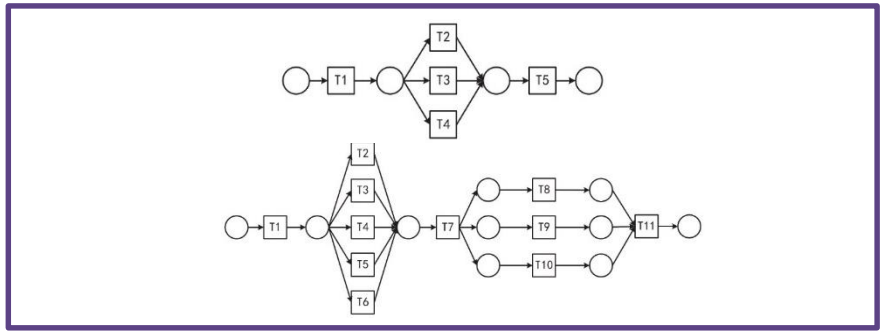
* BeehiveZ is an open source Business Process Model and Instance Management System, founded by ISE,Tsinghua University, https://code.google.com/p/beehivez/

Content

Feature Selection

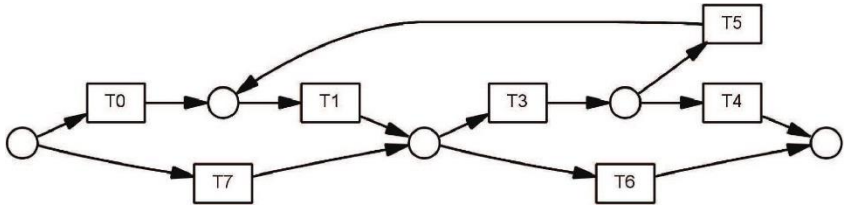
- **48 features are reduced to six features**
- **Two Criteria**
  - 1. The features which characterize a model's size should be removed.
    - *Number of node, Diameter, …*
  - 2. The features which characterize a model's connectors should be removed.
    - *Number of Connectors, Connector Heterogeneity, …*

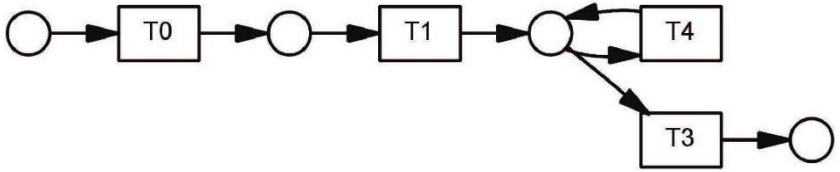


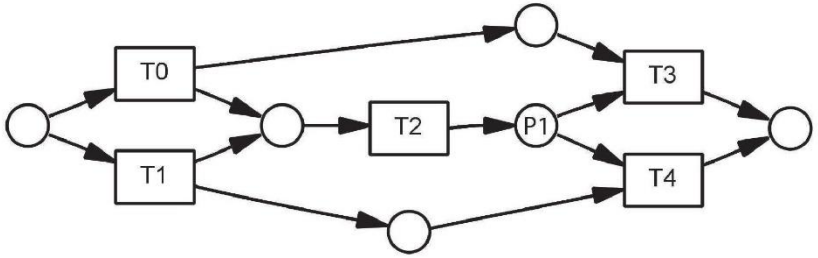


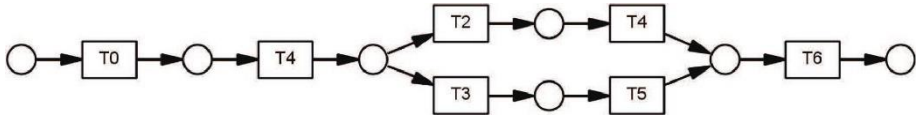
Feature Selection

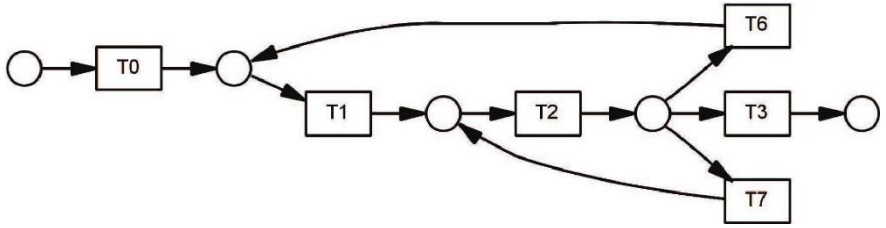
- Six Selected Feature
  - 1. Number of *invisible task* :
  - 2. Number of *duplicate task* :
  - 3. Number of *non-free choice* :
  - 4. Number of *arbitrary loop* :
  - 5. Number of *short loop* :
  - 6. Number of *nested loop* :

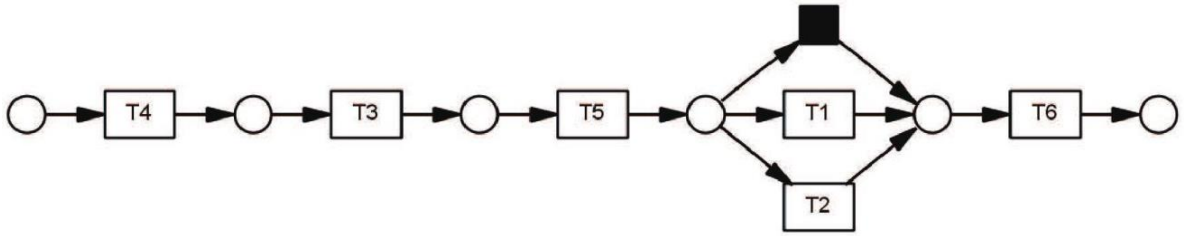










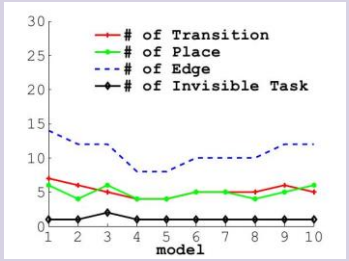
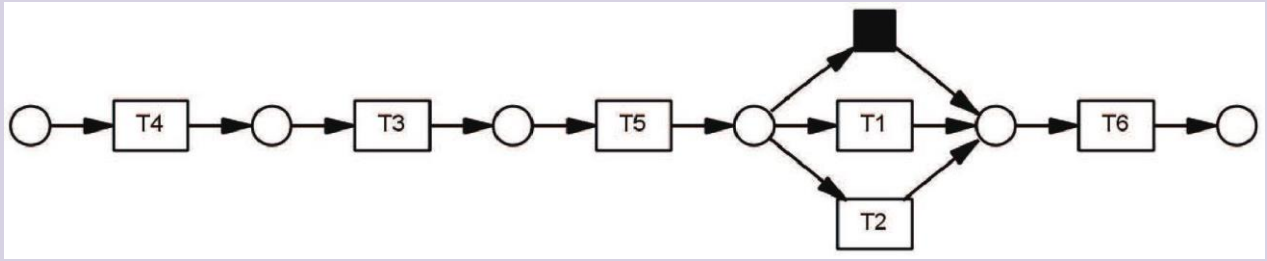
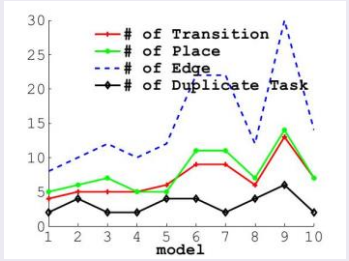
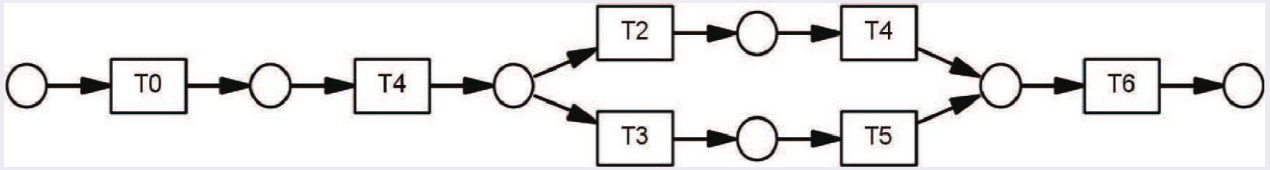
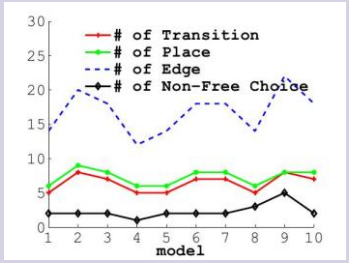
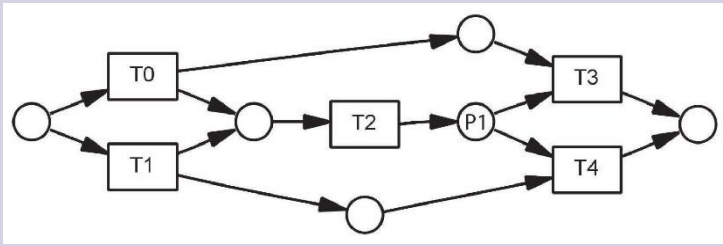


Content

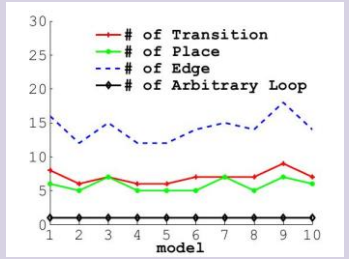
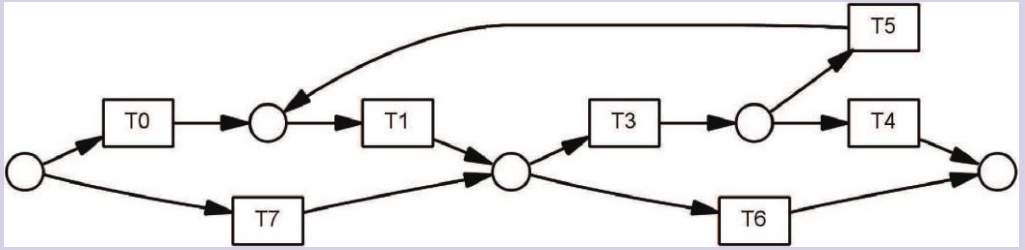
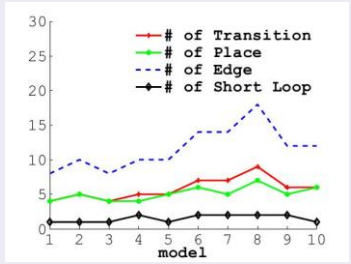
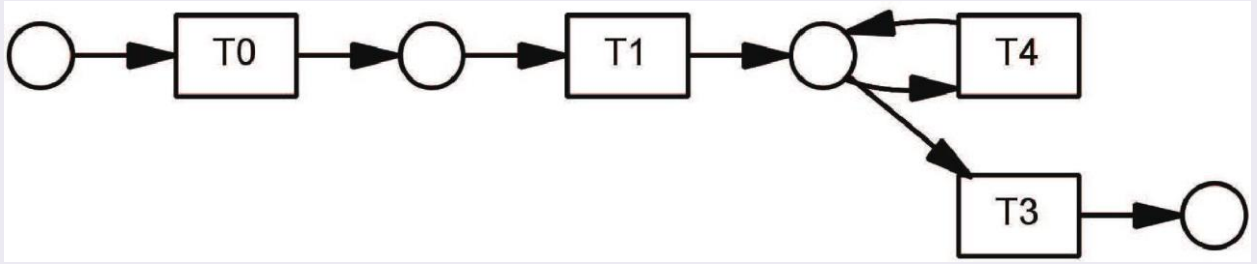
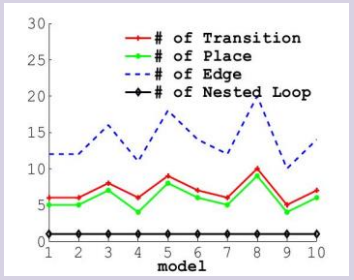
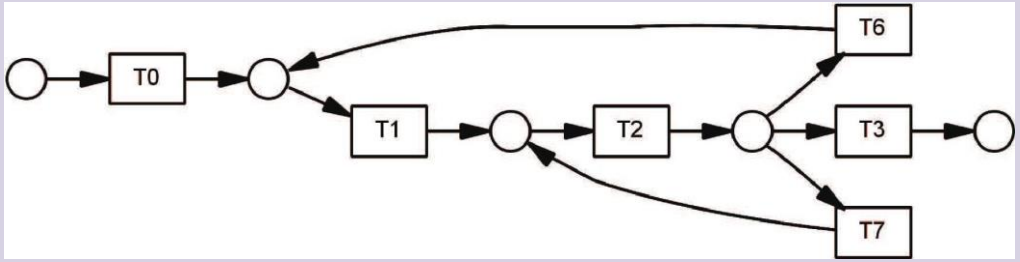
Model Set

- Inspired by the selected 6 process model features that characterize the model rediscoverability
- Artificially construct 10 process models for each process model feature
- Total 60 process models

Model Set

Feature	Statistic	Example
Invisible Task		
Duplicate Task		
Non-Free Choice		

Model Set

Feature	Statistic	Example
Arbitrary Loop		
Short Loop		
Nested Loop		

Content

Evaluation

- Data Set
  - Artificial Dataset
  - Boiler Manufacturer Dataset
  - High Speed Railway Dataset

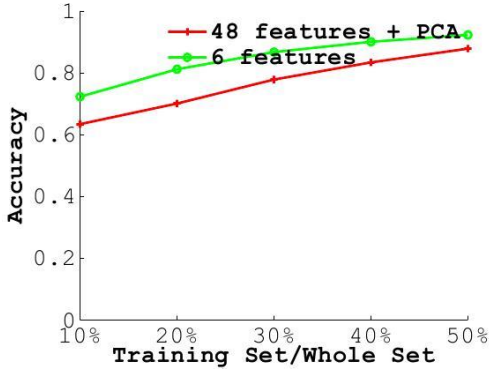
Table 1. Dataset Properties

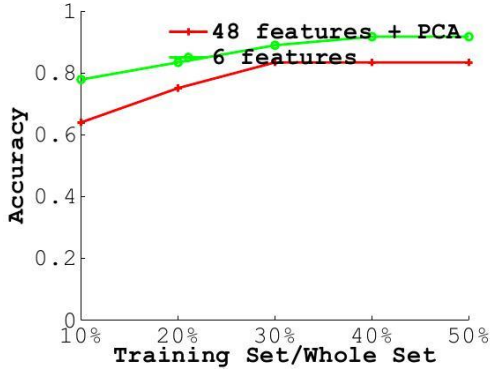
Dataset	Size	Average			Minimum			Maximum		
		#transitions	#places	#arcs	#transitions	#places	#arcs	#transitions	#places	#arcs
Artificial	270	6.100	6.244	13.233	2	3	4	13	14	30
Boiler	108	7.222	7.639	14.694	3	4	6	12	11	24
Trains	243	16.024	14.679	32.629	6	6	12	36	32	72

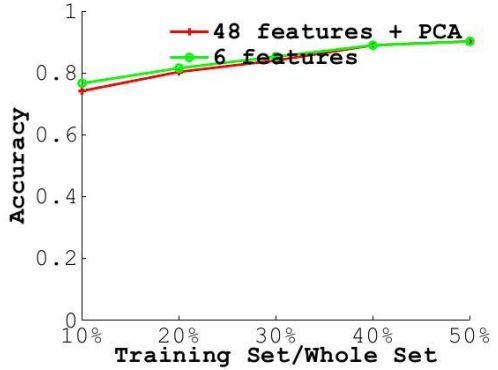
Evaluation on Feature Selection

- We select a fraction of process models (10%, 20%, 30%, 40%, and 50% respectively) from each datasets as training set to obtain the regression model, then recommend the most suitable process mining algorithm for the remaining models by applying the regression analysis

Accuracy

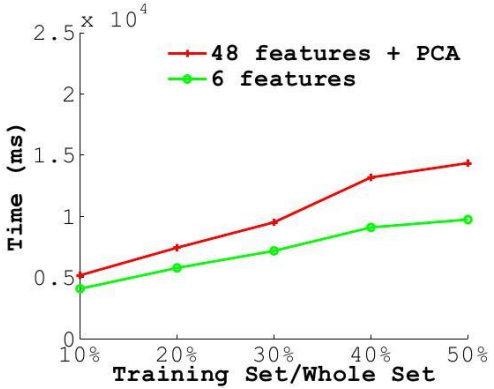


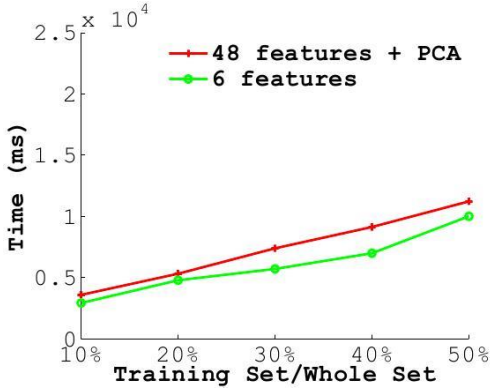




Artificial | Dongfang | Highspeed

Time





Evaluation on Universal Significant Reference Model Set

- We compare our proposed universal significant reference model set (we call it URM for short) with the reference models selected from the each datasets (We call it ORM for short). In the ORM, we select one third process models of each datasets as the reference models.

Table 2. Time Cost on Evaluating the Universal Significant Reference Model Set

Dataset	ORM(s)				URM(s)
	Mining	Training	Recommending	All	Recommending
Artificial	2789	83	14	2886	10
Boiler	1393	31	11	1435	10
Trans	18722	42	22	18786	15

Table 3. Accuracy on Evaluating the Universal Significant Reference Model Set

Dataset	Size	ORM		URM	
		#Correct	Accuracy	#Correct	Accuracy
Artificial	180	158	87.78%	166	92.22%
Boiler	72	60	83.33%	66	91.67%
Trains	162	146	90.12%	154	95.06%

Content

Conclusion

- Contribution
  - A universal significant reference model set is proposed.
  - A small set of process model features that are specializing on model rediscoverability are selected.
- Future work
  - further analyzing the models in this universal significant reference model set and the features that characterizing the model's rediscoverability
  - we hope to design a new process mining algorithm with better performance on model rediscoverability.

Question


תודה
Dankie
Gracias
Спасибо
شكرًا
Merci
Takk
Köszönjük
Terima kasih
Grazie
Dziękujemy
Dėkojame
Ďakujeme
Vielen Dank
Paldies
Kiitos
Täname teid
谢谢
Thank You
Tak
感謝您
Obrigado
Teşekkür Ederiz
감사합니다
Σας Ευχαριστούμ
ขอบคุณ
Bedankt
Děkujeme vám
ありがとうございます
Tack