Agrupamento interativo aplicado à mineração de processos de negócio

Interactive Trace Clustering

Thais Rodrigues Neubauer

Orientadora: Profa. Dra. Sarajane Marques Peres

Co-orientador: Prof. Dr. Marcelo Fantinato

Março de 2020

EACH | USP
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo

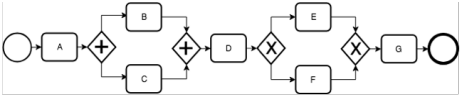
Agenda

- Mineração de processos
- Agrupamento interativo
- Experimentos
- Conclusões, limitações e trabalhos futuros

Mineração de processos

Modelo de processo

- Atividades e restrições.
- Conhecer processo real & sucesso na gestão.
- Muitas vezes, sem modelo.



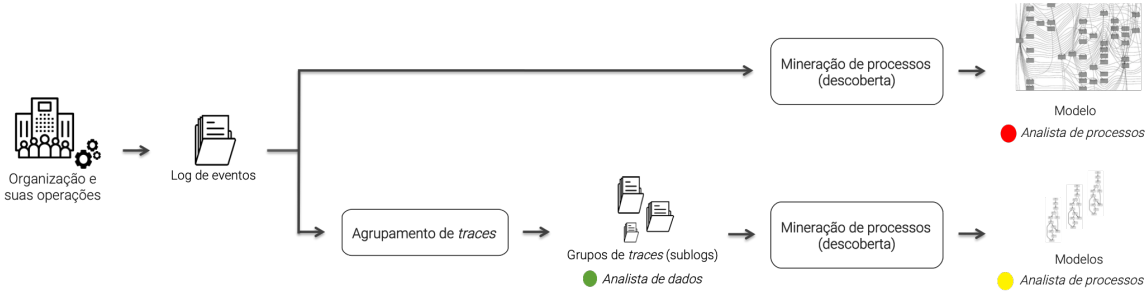
Mineração de processos

- Extrair conhecimento de logs de eventos.
- Descobrir, analisar e aprimorar processos.
- Conhecimentos de mineração de dados.

Caso	Evento
1	A
1	B
2	A
2	C

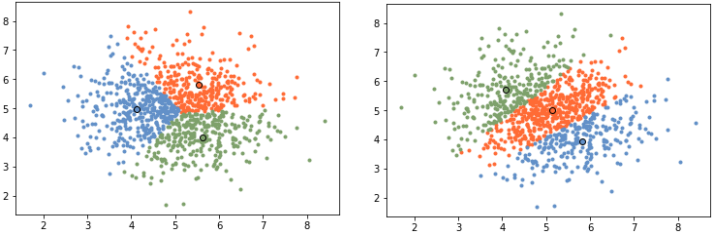
Trace	Eventos
1	A,B,C,D,E,G
2	A,C,B,D,E,G
3	A,B,C,D,F,G
4	A,C,B,D,F,G

Trace clustering



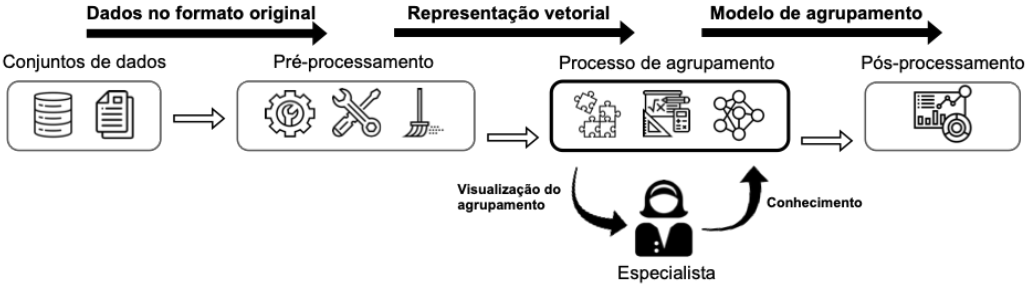
Agrupamento interativo

- **Agrupamento**: identificação de padrões com base em **similaridade**.
- Decisões como a função de similaridade e representação, algoritmo e outros hiperparâmetros.
- Ambiguidade e/ou inadequação de resultados.
- **Agrupamento interativo**: substituir/complementar com o conhecimento de especialista.



(a) Distância euclidiana (b) Similaridade cosseno

Agrupamento interativo



Agrupamento interativo

Tabela 1: Revisão de escopo - 50 estudos de agrupamento interativo

Etapa de coleta	**Etapa de inclusão**
Requisições split/merge	Aprendizado de medida de similaridade
Redistribuição de dados	Execução de split/merge
Restrições must-link/cannot-link	Aplicação de restrições must-link/cannot-link
Seleção de características	Seleção de semente
Ranqueamento de características	Realocacação de dado

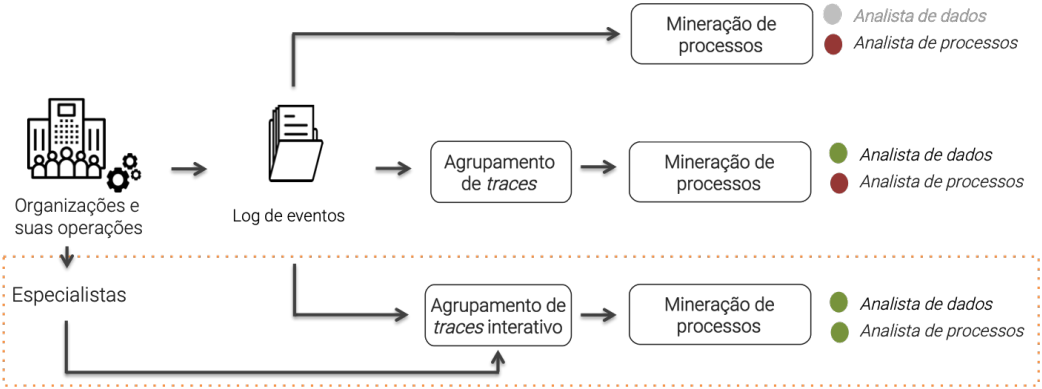
Agrupamento interativo: must-link/cannot-link com Cop-Kmeans

Algoritmo 1 COP-Kmeans(D, m, c)

$C_1 ... C_k$: centroides iniciais dos k grupos.
Enquanto não há convergência **faça**:
  Para $i \leftarrow 0$ **até** n **faça**:
    Associe o dado d_i ao grupo C_j mais próximo que não viole m e n.
    Se C_j não existir **então**
   Retorne $\{\}$
  Para $i \leftarrow 0$ **até** k **faça**:
    Atualize o centroide do grupo C_i.
Retorne $\{C_1 ... C_k\}$

Fonte: Adaptado de Wagstaff et al. (2001)

Problema de pesquisa e proposta



- *Interactive Trace Clustering*: Aplicação de *interactive clustering* em *trace clustering*.

Agrupamento interativo aplicado à mineração de processos de negócio

Trabalhos correlatos

- Aplicações envolvendo o especialista.
- O que consideramos como agrupamento interativo.
- Koninck et al. (2017) - classificado como semi-supervisionado.
  - definição de perfis por especialistas
  - *consensus clustering* ou *seeding* com reagrupamento.

Agrupamento interativo aplicado à mineração de processos de negócio

Abordagem de aplicação de agrupamento interativo desenvolvida

- Visualização: caracterização de grupos baseada na frequência dos valores de características de interesse.
- Portanto, conhecimento coletado baseado nos valores de características.
- Problema: mesmo valor de característica em **N** dados & restrições *must-link/cannot-link* entre **pares**.
- **Cop-Kmeans com Mapeamento de Conhecimento em Restrições (CMCR)**.

Procedimento para criação da estrutura de restrições *must-link*

$D_n = \{d_1 ... d_n\}$: conjunto D_n de índices dos n dados.
M : lista de restrições *must-link*.
Para $i \leftarrow 1$ **até** n **faça**:
  $M[i] = \{d_1 ... d_n\} - \{d_i\}$.
Retorne M

$$D_n \sim D_n$$

Procedimento para criação da estrutura de restrições *cannot-link*

$D1_{n1} = \{d1_1 ... d1_{n1}\}$: conjunto $D1_{n1}$ de índices dos $n1$ dados selecionados.
$D2_{n2} = \{d2_1 ... d2_{n2}\}$: conjunto $D2_{n2}$ de índices dos $n2$ dados selecionados.
N : lista de restrições *cannot-link*.
Para $i \leftarrow 1$ **até** $n1$ **faça**:
  Para $j \leftarrow 1$ **até** $n2$ **faça**:
    $N[i] = N[i] + \{d2_j\}$
    $N[j] = N[j] + \{d1_i\}$
Retorne N

$$D1_{n1} \bowtie D2_{n2}$$

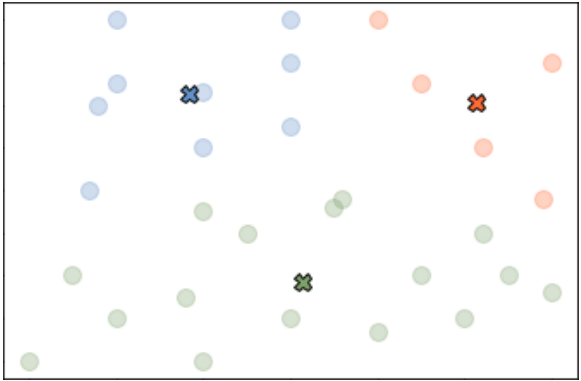
Agrupamento interativo aplicado à mineração de processos de negócio

Escolha de parte dos dados para restrições

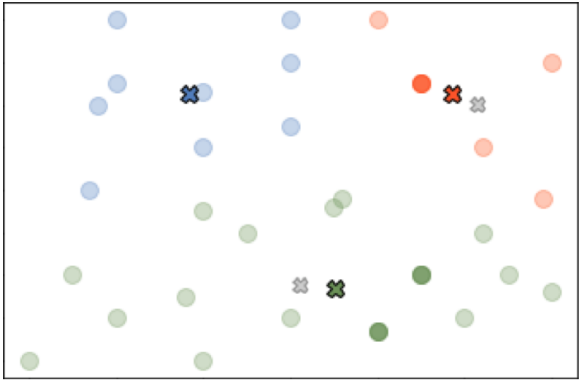
- Quantidade de restrições geradas é exponencial em relação à quantidade de dados selecionados.
- Mesmos resultados podem ser alcançados com parte dos dados?
- Escolha de dados únicos (espaço) *versus* escolha de repetições (força).

Case	Atividades							
	a_1	b_1	c_1	d_1	e_1	f_1	g_1	...
1	1	0	1	1	0	1	0	0
2	1	0	1	1	0	1	0	0
3	1	0	1	1	0	1	0	0
4	1	0	1	1	0	1	0	0
5	0	1	0	1	1	0	1	0
6	0	1	0	1	1	0	1	0

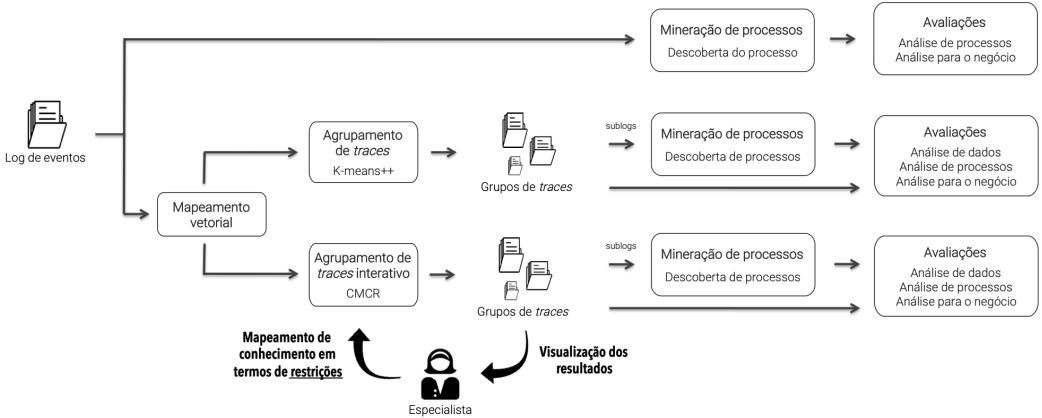
Efeito de repetições no agrupamento



Efeito de repetições no agrupamento



Experimentos



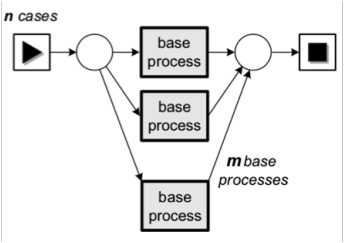
Experimentos - log de eventos sintéticos

Objetivo do experimento

Ambiente de teste independente de contexto, controlado.
Abstração de questões semânticas.

Representações

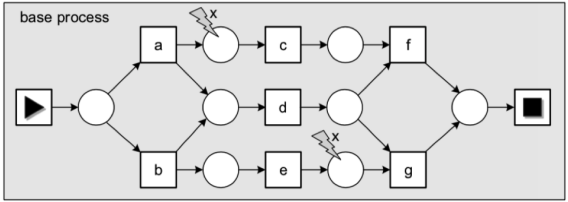
Perfis: {por atividade, por transição}
Modelo de contagem: {binária, tfidf}



* Baseado no conjunto "Benchmarking logs to test scalability of process discovery algorithm" (*4TU.ResearchData*, W. M. P. Van der Aalst, 2017)

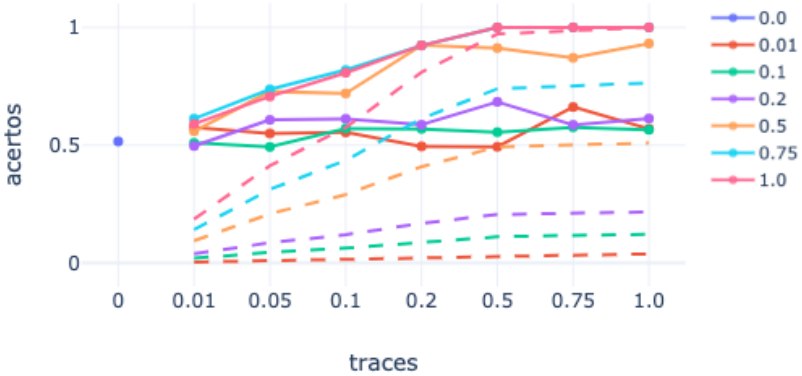
Experimentos - log de eventos sintéticos

Cenário	Especificação do log	#Grupos	Expectativa	Avaliações
0	$m = 100,\ n = 10.000$	$k = 100$	Separar réplicas.	Expectativa especialista.
1	$m = 100,\ n = 10.000$	$k = 2$	XOR-split.	Expectativa especialista, qualidade agrupamento, modelo de processo.
2	$m = 100,\ n = 10.000$	$k = 3$	XOR-split e ruído.	Expectativa especialista, qualidade agrupamento, modelo de processo.
3	$m = 5,\ n = 493$	$k = 3$	Réplicas $1 \sim 2$; $3 \bowtie 4$.	Expectativa especialista, qualidade agrupamento, modelo de processo.



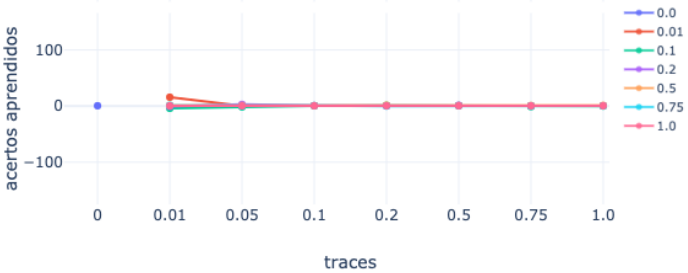
Experimentos - log de eventos sintéticos

- **Cenário 1** - *XOR-split*



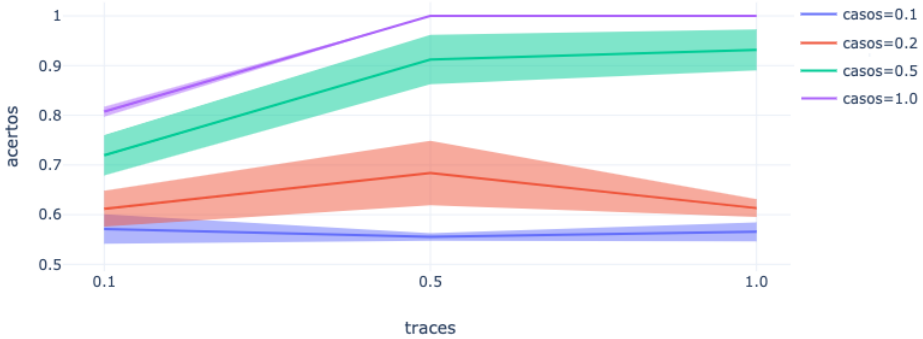
Experimentos - log de eventos sintéticos

- **Cenário 1** - *XOR-split*



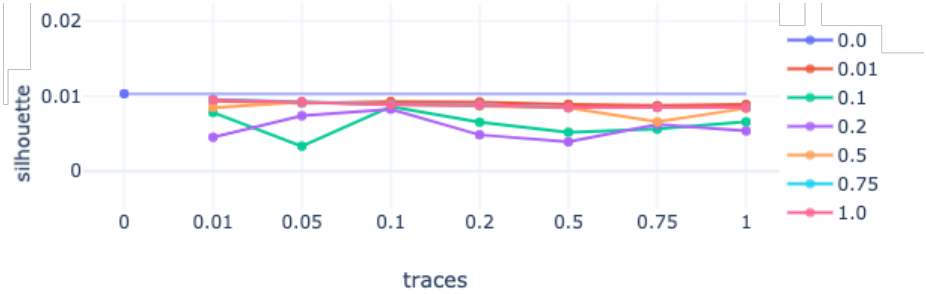
Experimentos - log de eventos sintéticos

- **Cenário 1** - *XOR-split*



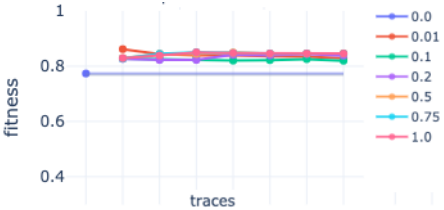
Experimentos - log de eventos sintéticos

- **Cenário 1** - *XOR-split*

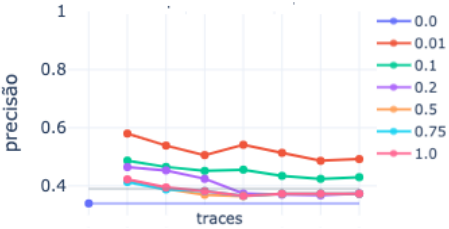


Experimentos - log de eventos sintéticos

- **Cenário 1** - *XOR-split*



(a) Fitness

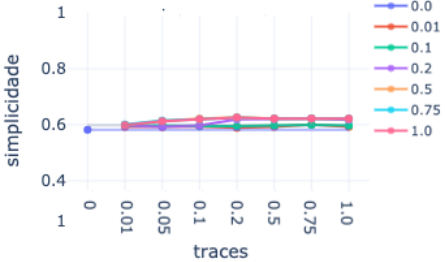


(b) Precisão

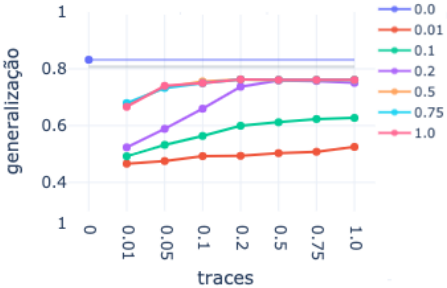
- Todos ao menos semelhantes ao log completo.
- Kmeans++: valores superiores com perfil por transição (acentuada na representação binária).
- Precisão: maiores variações entre as representações e *#traces* e casos.

Experimentos - log de eventos sintéticos

- **Cenário 1** - *XOR-split*



(c) Simplicidade



(d) Generalização

- Simplicidade: ao menos semelhantes ao log completo e ao Kmeans++.
- Generalização: abaixo do log completo e do Kmeans++; maior variação por *#traces* e caso.

Experimento - log de eventos reais

O log de eventos reais

- Relata eventos de um processo de gerenciamento de incidentes extraído de uma instância da plataforma *ServiceNow*^{TM} utilizada por uma empresa de TI.
- Composto por 24.918 incidentes, extraídos de 03/2016 – 02/2017.
- **Tarefa de mineração de processos:** predição de tempo de resolução de incidentes.
- *Trace clustering* → grupos com menor variação de tempo de resolução de incidentes → maior precisão.

	number	incident_state	active	reassignment_count	...	close_code	resolved_by	resolved_at	closed_at
0	INC0000045	New	True	0	...	code 5a	Resolved by 149	29/2/2016 11:29	2016-03-05 12:00:00
1	INC0000045	Resolved	True	0	...	code 5a	Resolved by 149	29/2/2016 11:29	2016-03-05 12:00:00
2	INC0000045	Resolved	True	0	...	code 5a	Resolved by 149	29/2/2016 11:29	2016-03-05 12:00:00

Experimento - log de eventos reais

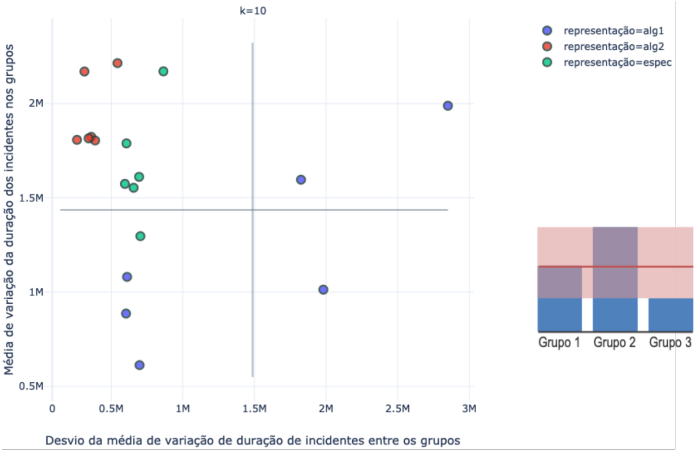
Escolha da representação vetorial

Conjunto de atributos: {alg1, alg2, espec}
Composição de atividade: {por atributo individual, por combinação de atributos}
Modelo de contagem: {binária, tf, tfidf}
Valores de k: {3,5,10}
Algoritmo e métrica de similaridade: Kmeans++ e distância euclidiana.

Avaliações utilizadas na escolha

- Semelhança do tempo de resolução dos incidentes de cada grupo - desvio padrão .
- Qualidade do agrupamento do ponto de vista de mineração de dados - índice Silhouette.
- Questionário aplicado junto a 5 especialistas de mercado.

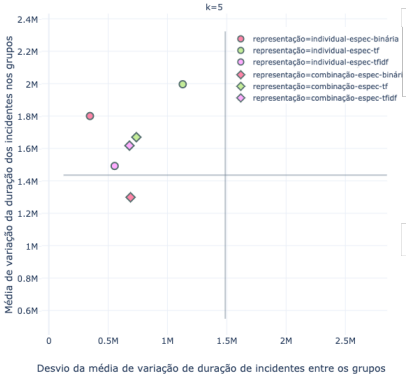
Experimento - log de eventos reais



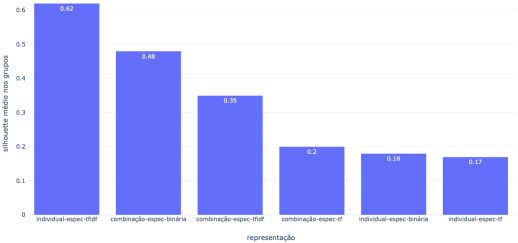
* Silhouette: combinação com representação binária; individual com representação tfidf.

Experimento - log de eventos reais

- Seleção final da representação:



(a)



(b)

Experimento - log de eventos reais

CMCR com o log real

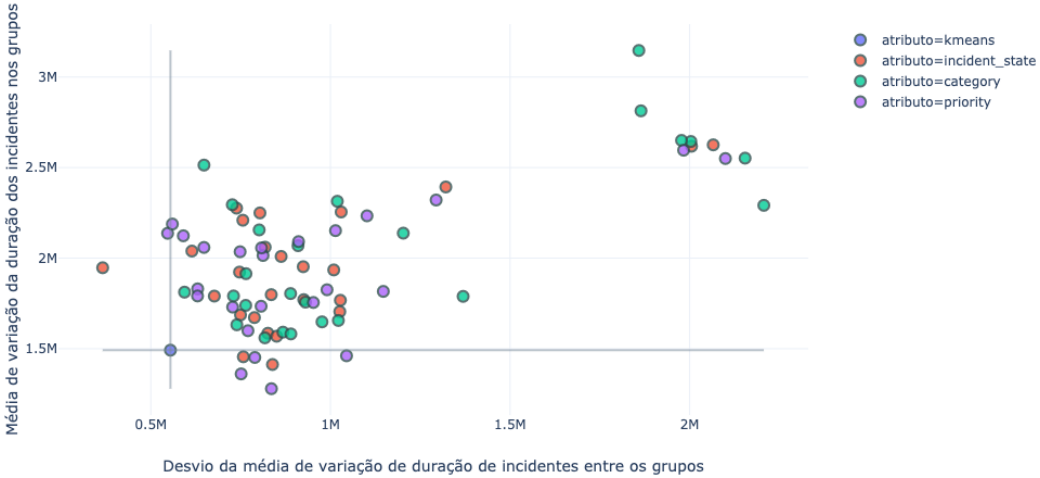
- Aplicação do segundo questionário para elucidar restrições (2 acadêmicos, 2 mercado + 3 controle).
- Selecionados 3 conjuntos de restrições: uma para cada atributo.

incident_state:	incidentes que passam por estados "Awaiting ..." no mesmo grupo (*must-links*).
priority:	separar prioridades {1,2}, {3}, {4} (*cannot-links*).
category:	incidentes de categorias de rápida resolução no mesmo grupo (*must-links*).

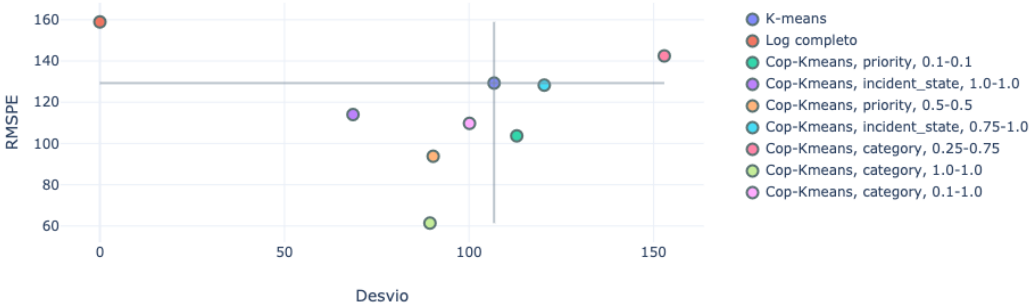
Avaliações utilizadas na escolha

- Média de variação da duração dos incidentes nos grupos.
- *Root Mean Square Percentage Error* (RMSPE) ou raiz do erro quadrático médio: a partir dos resultados de um preditor baseado em Sistemas de Transição Anotado (STA).
- Qualidade de agrupamento - Silhouette.
- Métricas de modelos de processo.

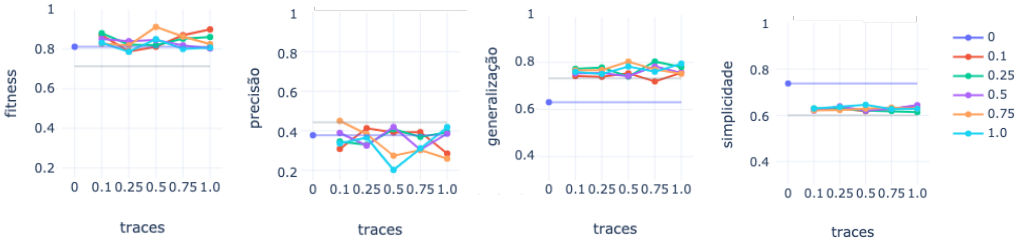
Experimento - log de eventos reais



Experimento - log de eventos reais



Experimento - log de eventos reais



Conclusão

1. Abordagem CMCR proposta é capaz de alcançar o objetivo (expectativa) do especialista.
   - Utilização de restrições parciais não garante o alcance do objetivo do especialista.
   - A inclusão de restrições não prejudica o aprendizado dos dados não envolvidos.
2. Qualidade do agrupamento (Silhouette) com CMCR:
   - Log de eventos sintéticos: sem degradação significativa.
   - Log de eventos reais: degradação mais significativa em algumas execuções.
3. Qualidade de modelo de processos com CMCR: perda de precisão.
   - Log de eventos sintéticos: perda de generalização (mais significativa).
4. Representações (log de eventos sintéticos):
   - Qualidade do agrupamento: perfil por atividades um pouco melhor;
   - Qualidade de modelo de processo: perfil por transição um pouco melhor com Kmeans++.
5. Tarefa de mineração de processos (log de eventos reais): melhora da predição de tempo.
6. Esforço do especialista (log de eventos reais): 1h/interação - baixo se interesse do especialista.

Limitações e oportunidades

- Aprimorar experimentos para melhor explorar os efeitos de aleatoriedade.
- Aumentar a quantidade de especialistas participantes.
  - Tempo disponível/engajamento dos especialistas.
  - Concordância e interessabilidade.
- Medidas de esforço do especialista (subjetividade).
- Medidas de qualidade (subjetividade).

Agradecimento

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Código de Financiamento 001.

Agrupamento interativo aplicado à mineração de processos de negócio

Interactive Trace Clustering

Thais Rodrigues Neubauer

Orientadora: Profa. Dra. Sarajane Marques Peres

Co-orientador: Prof. Dr. Marcelo Fantinato

Março de 2020

EACH | USP campus capital
Escola de Artes, Ciências e Humanidades
Universidade de São Paulo