ALEX MEINCHEIM

UMA CONTRIBUIÇÃO AO ESTUDO DE MEDIDAS DE SIMILARIDADE APLICADAS NO AGRUPAMENTO INCREMENTAL DE INSTÂNCIAS DE PROCESSOS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática.

Curitiba
2018

ALEX MEINCHEIM

UMA CONTRIBUIÇÃO AO ESTUDO DE MEDIDAS DE SIMILARIDADE APLICADAS NO AGRUPAMENTO INCREMENTAL DE INSTÂNCIAS DE PROCESSOS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica do Paraná como requisito parcial para obtenção do título de Mestre em Informática.

Área de concentração: Ciência da Computação

Orientador: Prof. Dr. Edson Emílio Scalabrin

Curitiba
2018

Dados da Catalogação na Publicação
Pontifícia Universidade Católica do Paraná
Sistema Integrado de Bibliotecas – SIBI/PUCPR

M514c 2018

Meincheim, Alex
Uma contribuição ao estudo de medidas de similaridade aplicadas no agrupamento incremental de instâncias de processos / Alex Meincheim; orientador, Edson Emílio Scalabrin. -- 2018
100 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Paraná, Curitiba, 2018
Bibliografia: f.94-100

1. Informática. 2. Mineração de dados (Computação). 3. Algoritmos. 4. Controle de processo. 5. Administração da produção. 6. Desenvolvimento organizacional. I. Scalabrin, Edson Emílio. II. Pontifícia Universidade Católica do Paraná. Programa de Pós-Graduação em Informática. III. Título.

CDD 20. ed. – 004.068

Biblioteca Central

Pontifícia Universidade Católica do Paraná

PUCPR

GRUPO MARISTA

ATA DE SESSÃO PÙBLICA

DEFESA DE DISSERTAÇÃO DE MESTRADO Nº 01/2018

PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA – PPGIa
PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ - PUCPR

Em sessão pública realizada às 14h00 de **23 de Fevereiro de 2018**, no Auditório Guglielmo Marconi – Bloco 8, ocorreu a defesa da dissertação de mestrado intitulada **"Uma Contribuição ao Estudo de Medidas de Similaridade Aplicadas no Agrupamento Incremental de Instâncias de Processos"** apresentada pelo aluno **Alex Meinchem** , como requisito parcial para a obtenção do título de **Mestre em Informática**, na área de concentração **Ciência da Computação,** perante a banca examinadora composta pelos seguintes membros:

Prof. Dr. Edson Emílio Scalabrin - PUCPR

Prof. Dr. Bráulio Coelho Avila – PUCPR

Prof. Dr. Julio Cesar Nievola – PUCPR

Prof. Dr. Eduardo Alves Portela Santos – PUCPR/PPGEPS

Após a apresentação da dissertação pelo aluno e correspondente arguição, a banca examinadora emitiu o seguinte parecer sobre a tese:

Membro	**Parecer**
Prof. Dr. Edson Emílio Scalabrin	(X) Aprovado ( ) Reprovado
Prof. Dr. Bráulio Coelho Avila	(X) Aprovado ( ) Reprovado
Prof. Dr. Julio Cesar Nievola	(X) Aprovado ( ) Reprovado
Prof. Dr. Eduardo Alves Portela Santos	(X) Aprovado ( ) Reprovado

Portanto, conforme as normas regimentais do PPGIa e da PUCPR, a tese foi considerada:

(X) **APROVADO**

(aprovação condicionada ao atendimento integral das correções e melhorias recomendadas pela banca examinadora, conforme anexo, dentro do prazo regimental)

( ) **REPROVADO**

E, para constar, lavrou-se a presente ata que vai assinada por todos os membros da banca examinadora. Curitiba, 23 de Fevereiro de 2018.

Prof. Dr. Edson Emílio Scalabrin

Prof. Dr. Bráulio Coelho Avila

Prof. Dr. Julio Cesar Nievola

Eduardo Alves Portela Santos

DEDICATÓRIAS

Dedico este trabalho aos meus pais, Célio e Odila, pelo apoio incondicional e a minha noiva Rosani pelo carinho e incentivo em todos os momentos.

AGRADECIMENTOS

Agradeço primeiramente a Deus, por me guiar, iluminar e me socorrer nos momentos difíceis.

Aos meus pais Célio e Odila pela educação base para minha vida e pelo apoio incondicional na minha trajetória.

A minha querida noiva Rosani por todo o carinho, amor, parceria e compreensão nos momentos de ausência.

Ao meu orientador Edson Emílio Scalabrin, pelos ensinamentos, esclarecimentos e toda a atenção empreendida. De fato, foram cruciais para o desenvolvimento deste trabalho.

Ao meu amigo e companheiro de estudos, Cleiton dos Santos Garcia, por todas as dicas e valiosas discussões.

Aos professores Júlio Cesar Nievola, Bráulio Coelho Ávila, Eduardo Alves Portela Santos e Andreia Malucelli que contribuíram para a realização deste trabalho.

Aos colegas da WEG, em especial, Paulo Sérgio dos Santos, pelo apoio concedido nesta trajetória.

À PUCPR, ao Programa de Pós-Graduação em Informática (PPGIa) e a Fundação CAPES pela oportunidade e suporte oferecido.

Por fim, a todos os amigos que direta ou indiretamente contribuíram para a execução deste trabalho.

RESUMO

Em organizações contemporâneas, os processos—cada vez mais flexíveis—são elaborados com o objetivo de atender a dinâmica demanda do mercado, que por sua vez permitem amplo espectro de comportamentos e decisões em tempo de execução. A Mineração de Processos aplicada nestas organizações geralmente resulta em modelos de processos difíceis de compreender. Uma forma de melhorar o resultado da Mineração de Processos, como etapa preliminar, é agrupar, por meio de uma medida de similaridade, as instâncias de processos. Tal agrupamento deve: (1) melhorar a acurácia e compreensão dos modelos descobertos; e (2) facilitar a descoberta de variantes e desvios de processos. A hipótese é que a medida de similaridade empregada impacta diretamente no resultado do agrupamento de instâncias de processos. Neste contexto, foi realizado um estudo comparativo para avaliar o impacto das medidas de similaridade de modelos de processos na qualidade dos agrupamentos de instâncias de processos complexos; as medidas de similaridades experimentadas foram: correspondência de rótulos, distância de grafos e análise da causalidade das atividades. Os *logs* de eventos utilizados são de processos semiestruturados e não estruturados, compostos por base de dados sintéticos e reais. A partir dos resultados obtidos concluiu-se que as medidas de similaridade empregadas na tarefa de agrupamento de instâncias de processos impactam diretamente na qualidade dos modelos descobertos de processos. Verificou-se, em particular, que para os logs de eventos reais—processos não estruturados—, a medida de similaridade que considera as relações de transições adjacentes apresentou resultados superiores as demais medidas; embora não tenha ocorrido diferença estatística em todos os *log*s de eventos considerados. Aqui, a principal contribuição está ligada a hipótese de que as medidas de similaridade impactam diretamente na qualidade dos agrupamentos de instâncias de processos. Essa verificação sugere medidas de similaridade mais apropriadas para a tarefa de agrupamento quando aplicado em conjunto com a Mineração de Processos em processos não estruturados.

Palavras-chaves: Mineração de Processos, Agrupamento de Instâncias de Processos, Medidas de Similaridade, Modelos de Processos.

ABSTRACT

In contemporary organizations, processes—increasingly flexible—are designed to attend the dynamic market demand, which in turn allows widely spectrum of behavior and decisions at runtime. Process Mining applied in these organizations usually results in processes models difficult comprehension. One way to improve the outcome of Process Mining, as a preliminary step, is to cluster through a similarity measure the process instances. This clustering should: (1) improve the accuracy and comprehension of discovered process models; and (2) facilitating the discovery of variants and deviations of processes. The similarity measure used directly impacts on the result of the process instances clustering. In this context, a comparative study was carried out to evaluate the impact of similarity measures of process models on the quality of the clustering in instances of unstructured processes. The similarity measure experienced were: label matching, graph edit distance, and causal dependencies between activities. The event logs used represents semi-structured and unstructured processes, composed of the synthetic and real-life database. From the results obtained it was concluded that the similarity measures used in the clustering task of process instances directly affect the quality of the discovered process models. For real-life event logs that represent unstructured processes, it was verified that the similarity measure that considers the relations of adjacent transitions presented better results when compared to the other measures; although there was not a statistical difference in all the event logs considered. The main contribution of this work refers to the hypothesis that similarity measures directly impact in the clustering quality of process instances. This verification suggests similarity measures more appropriated for the clustering task when applied in conjunction with the Process Mining techniques in unstructured processes.

Keywords: Process mining, Process instance clustering, Similarity measures, Process models

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

PM	Modelo de Processos
BPMN	Notação de Modelagem de Processos de Negócio
CPI	Melhoria Contínua de Processos
TQM	Gestão da Qualidade Total
BPM	Gerenciamento de Processos de Negócio
BPA	Análise de Processos de Negócio
BAM	Monitoramento de Atividades de Negócio
ProM	Framework de Mineração de Processos

CAPÍTULO 1 - INTRODUÇÃO

Em organizações contemporâneas, os processos—cada vez mais flexíveis—são criados com o objetivo de atender a dinâmica demanda do mercado, que por sua vez permitem um amplo espectro de comportamentos e pontos de decisões [1]. O processo de melhoria de processos que compreende o seu mapeamento, entendimento e controle é um grande desafio, na medida em que ele dispende muito esforço quando aplicado em processos flexíveis devido as variações de comportamento. Técnicas tradicionais de mapeamento de processos são muito utilizadas para entender como os processos ocorrem na prática, contudo, o resultado na grande maioria são modelos de processos que ocultam ou generalizam comportamentos, não refletindo a realidade do processo investigado.

A Mineração de Processos permite a observação da execução do processo com base em dados de eventos reais, e propõe métodos e ferramentas para fornecer agilidade aos diagnósticos [6], reduzindo as lacunas entre a prática e os modelos conceituais. A Mineração de Processos pode ser utilizada para: (1) descobrir e gerar modelos de processos reais sem conhecimento *a priori*; (2) verificar a conformidade de um modelo de processo teórico por meio de um modelo descoberto; (3) aplicar e estender um modelo de processo por meio de informações de desempenho, gargalos, recursos e custos [6] com o objetivo de trazer novas perspectivas de análises.

Neste contexto, técnicas de descoberta de processos ([17] [19] [21] [23] [24] [25] [58]) possuem problemas para lidar com *log* de eventos provenientes de processos flexíveis, pois o resultado geralmente são modelos de processos de difícil compreensão devido ao elevado número de atividades e transições, conhecidos também como modelos espaguete. Uma forma de melhorar os resultados da Mineração de Processos é aplicar técnicas para agrupar as instâncias de processos presente no *log* de eventos como etapa anterior da descoberta com objetivo de

melhorar a acurácia e compreensão dos modelos ([1] [29]) e descobrir variantes e desvios do processo analisado [5].

Como parte do processo de agrupamento, a medida de similaridade empregada impacta diretamente no resultado dos modelos descobertos de processos [37]. Diversas técnicas foram propostas com o objetivo de medir a similaridade entre dois modelos, sendo elas segregadas em quatro grupos de acordo com as técnicas empregadas [3]: (1) Medidas baseadas na correspondência entre nós e arestas; (2) Medidas baseadas na distância de edição de grafos; (3) Medidas baseadas na dependência causal entre atividades; e (4) Medidas baseadas na comparação de conjuntos de traços ou *logs* de eventos.

Neste contexto, este trabalho propõe a realização do estudo de diferentes medidas de similaridade de processos e avaliar como elas impactam na qualidade do agrupamento de instâncias de processos complexos—i.e., semiestruturados ou não-estruturados. Para realizar este estudo foi conduzido um experimento envolvendo diferentes medidas de similaridade aplicadas no processo de agrupamento de instâncias de processos em *log* de eventos sintéticos e reais. A tarefa de descoberta de modelos de processo, a partir de um conjunto de instâncias de processos, foi realizada pelo minerador *Inductive Miner*. Assim, nesta ordem, realizada a tarefa de agrupamento de instância de processos, cada grupo foi submetido ao minerador para descoberta do modelo de processos que o representava. Posteriormente, os modelos de processos foram avaliados sob métricas de qualidade de modelos de processos como o *Recall, Precision* e *F1 Score*.

1.1 Problema de Pesquisa

Técnicas tradicionais de Mineração de Processos possuem limitações para lidar com grandes *logs* de eventos oriundos de processos não estruturados, de modo que os resultados obtidos são frequentemente complicados e difíceis de entender [2]. Com o objetivo de extrair modelos de processos compreensíveis, técnicas de agrupamento têm sido aplicadas como etapa preliminar a Mineração de Processos para melhorar a compreensão dos modelos ([1] [29]), e descobrir variantes e desvios de processos [5].

A medida de similaridade empregada no processo de agrupamento impacta diretamente no resultado dos modelos de processos resultantes [37]. Diferentes métricas de similaridade entre modelos foram propostas, aplicadas por exemplo, no gerenciamento de repositórios de processos de negócio para realizar buscas de processos nas bases de dados, na avaliação de conformidade, e na descoberta de serviços [3].

Desta forma, considerando que a escolha da medida de similaridade impacta diretamente no resultado do agrupamento e consequentemente no resultado obtido da Mineração de Processos, o problema de pesquisa neste trabalho se coloca por meio das seguintes questões: Diferentes técnicas que medem a similaridade entre modelos de processos impactam na qualidade do agrupamento de instâncias, e consequentemente no resultado obtido da Mineração de Processos? É possível estabelecer medidas de similaridades mais apropriadas para o agrupamento de processos, em particular, quando aplicadas em conjunto com a Mineração de Processos?

1.2 Objetivos

O objetivo geral é analisar se as diferentes medidas de similaridade de modelos de processos impactam diretamente na qualidade do agrupamento de instâncias de processos. Tal análise visa estabelecer medidas de similaridade mais apropriadas para o agrupamento de instâncias de processos, quando aplicadas em conjunto com a Mineração de Processos. Para alcançar esse objetivo, os seguintes objetivos específicos se fazem necessários:

- Selecionar diferentes medidas de similaridade de modelos de processos por meio de propriedades de medidas de similaridade voltadas a modelos de processos;
- Construir um ambiente para avaliação das diferentes medidas de similaridade de modelos de processos aplicáveis ao agrupamento de instâncias de processos;

- Analisar os resultados quanto a qualidade dos modelos descobertos de processos gerados a partir dos agrupamentos de instâncias de processos com as diferentes medidas de similaridade de modelos de processos.

1.3 Motivação

A principal motivação diz respeito em facilitar, por meio de uma etapa preliminar de agrupamento de instâncias de processos, a descoberta e análise de modelos de processos a partir de *log* de eventos de processos flexíveis complexos; i.e., log de eventos de processos que podem conter muitas instâncias de processos distintas. Diversas abordagens de agrupamentos de instâncias foram propostas para diminuir a complexidade dos modelos gerados, e identificar variantes e desvios de processos [4]. As medidas de similaridade utilizadas em conjunto com as técnicas de agrupamento impactam diretamente no resultado dos agrupamentos de instâncias de processos, e consequentemente, quando aplicadas em conjunto com a Mineração de Processos, elas geram modelos que podem melhor descrever o comportamento exibido no *log* de eventos de processos flexíveis.

Assim, para a tarefa de agrupamento de instâncias de processos em ambientes flexíveis, a boa escolha de medidas similaridade de modelos de processos permite gerar modelos com maior qualidade; i.e., descrevendo cada processo com maior clareza e simplicidade.

1.4 Processo de trabalho

Com objetivo de estruturação do trabalho de pesquisa definiu-se um processo básico para condução do mesmo. Tal processo consistiu nas fases de preparação, estruturação, execução e análise dos resultados, a saber:

- Fase 1 – Preparação da pesquisa: realização de estudos sobre Mineração de Processos, agrupamento de instâncias de processos e medidas de similaridade de processos focadas em modelos de processos.

- Fase 2 – Estruturação da pesquisa: definição do método de pesquisa a ser empregado na execução do trabalho, incluindo etapas e objetivos claramente definidos. A experimentação foi o método adotado.
- Fase 3 – Execução da pesquisa: condução da pesquisa realizada conforme o seu plano experimental. Em seguida a execução da pesquisa experimental, realizou-se a análise dos resultados coletados com a validação ou não das hipóteses elaboradas.
- Fase 4 – Análise dos Resultados: discussão dos resultados obtidos na fase de execução da pesquisa, descrevendo as considerações finais, conclusões e trabalhos futuros.

1.5 Estrutura do documento da dissertação

A organização desta dissertação é apresentada a seguir:

- No Capítulo 1, foi apresentada a contextualização do problema de pesquisa, os objetivos gerais e específicos.
- No Capítulo 2, é apresentada uma visão geral dos conceitos relacionados a pesquisa em questão, contemplando processos de negócios, tipos de processos, mineração de dados e métricas de qualidade.
- No Capítulo 3, são apresentadas diferentes medidas de similaridade de modelos de processos.
- No Capítulo 4, é apresentado o método de pesquisa utilizado neste trabalho, as etapas e preparação do método.
- No Capítulo 5, é apresentada a análise e discussão dos resultados e limitações identificadas desta pesquisa.
- No Capítulo 6, são apresentadas as conclusões finais e trabalhos futuros.

1.6 Considerações sobre o Capítulo

Este capítulo apresentou uma contextualização sobre processos flexíveis, limitações das técnicas de Mineração de Processos quando aplicadas em *log* de eventos que

representam processos não estruturados, e como lidar com estes problemas de forma a facilitar as análises em processos não estruturados por meio do agrupamento de instâncias de processos. Em seguida foi apresentado o problema de pesquisa, abordando questões sobre medidas de similaridade aplicadas no agrupamento de instâncias de processos, motivação, objetivos gerais e específicos. Por último foi descrito o processo de trabalho.

CAPÍTULO 2 - REVISÃO DA LITERATURA

Este capítulo apresentará os conceitos necessários para o desenvolvimento deste trabalho. Serão abordados conceitos relacionados a modelos de processos, tipos de processos, mineração de processos, e técnicas de agrupamento de instâncias de processos.

2.1 Modelos de Processos

Um processo pode ser visto como uma ordenação específica de atividades de trabalho definidas no tempo e no espaço, contendo um ponto de início e um ponto de fim, assim como entradas e saídas claramente identificadas [28]. Processos podem ser representados por modelos. Esses últimos são amplamente utilizados como uma fonte de informações e procedimentos para auxiliar no gerenciamento da complexidade organizacional [6]. Diversas notações são utilizadas para representar um modelo de processo, como Redes de Petri, BPMN, YAML, EPC, sendo as Redes de Petri uma notação que representa graficamente modelos de processos de forma simples e intuitiva [18].

Uma rede de Petri é definida como um gráfico bipartido constituído por espaços e transições. Ela é uma estrutura estática, mas operada por regras de disparo de *tokens*, onde tais *tokens* podem fluir pela rede [17]. De maneira formal, uma rede de Petri é uma tripla $N = (P, T, F)$, onde:

- P representa um conjunto finito de espaços;
- T representa um conjunto de transições; e
- F representa um conjunto de arcos direcionados, chamados de relação de fluxo, onde $F \subseteq (P \times T) \cup (T \times P)$.

O estado de uma Rede de Petri é determinado pela distribuição dos *tokens* nos lugares referindo-se como marcação. Uma Rede de Petri marcada é representada por

um par de (N, M), onde $N = (P, T, F)$ é uma rede de Petri, e M é um conjunto de *tokens* sobre P que denotam a marcação da rede. A Figura 1 ilustra uma rede de Petri.

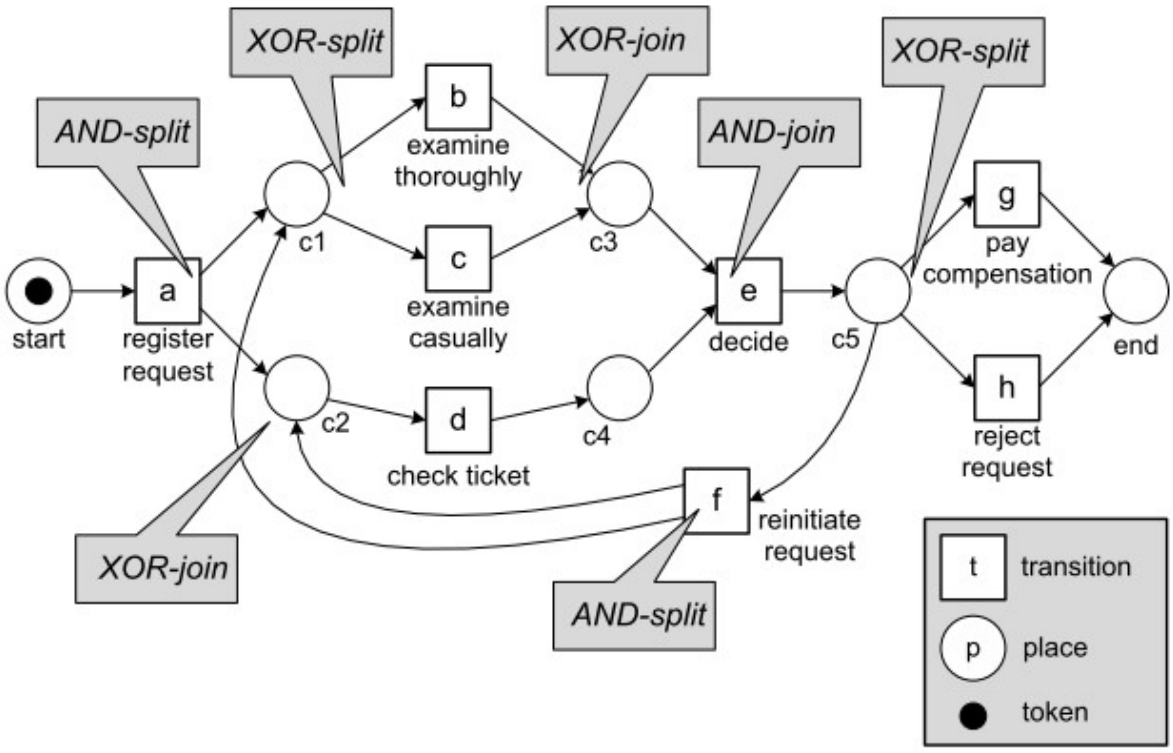


Figura 1- Petri net marcada. Fonte: [6].

O comportamento de uma Rede de Petri é definido por regras de disparos, onde uma transição é habilitada se cada um dos espaços contém um *token* na entrada. Quando uma transição é disparada, a mesma consome os *tokens* de cada espaço em sua entrada e produz um *token* em sua saída. Na Mineração de Processos, Redes de Petri são utilizadas em algoritmos de mineração como *Alpha miner* [17] e *Inductive Miner* [19] para representar os modelos de processos descobertos.

2.1.1 Variantes de Processo

A variabilidade de processos de negócio está relacionada ao gerenciamento flexível de processos, sendo possível distingui-los em três fases do ciclo de vida de modelos de processos customizáveis [27]:

- **Ciclo de projeto:** fase em que o modelo de processo personalizável é criado, de tal maneira que as decisões tomadas nesta fase afetarão toda a família de processos na fase seguinte.
- **Ciclo de customização:** nesta fase, o modelo de processo personalizável é customizado para realizar uma variante de processo particular. Cada customização é potencialmente geradora de variantes nas instâncias de um processo;
- **Ciclo de execução:** o modelo de processo customizado é colocado em prática para instâncias individuais de processos, de modo que, em cada instância, decisões são tomadas em tempo de execução. Deve-se notar que tais decisões afetam apenas uma única instância de processo.

Estes ciclos são contextualizados na Figura 2. Aqui, pode-se observar que um modelo personalizável permite a criação de um ou mais modelos customizados, que por sua vez podem gerar diferentes tipos de instâncias de processos.

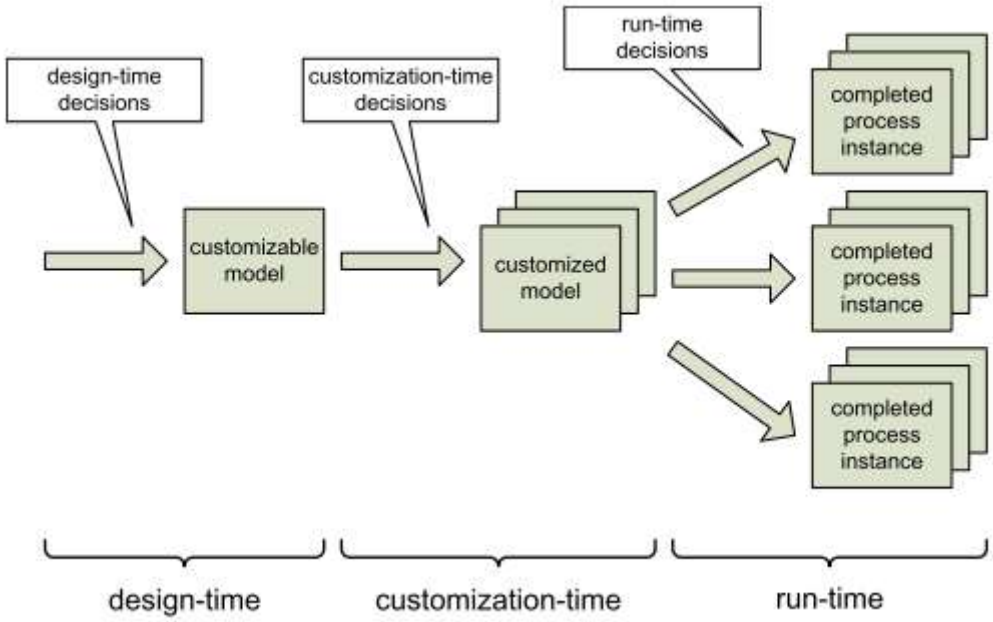


Figura 2 - Ciclo de vida de processos flexíveis. Fonte: [27].

Por meio da visão sobre o ciclo de vida de modelos de processos personalizáveis, a flexibilidade e a variabilidade estão relacionadas a diferentes estágios em que as decisões sobre um processo de negócio são tomadas [27]. A flexibilidade está relacionada as decisões tomadas em tempo de execução, enquanto que a variabilidade está relacionada as decisões tomadas nos ciclos de projeto de customização. Estas características afetam a natureza dos processos, tornando os processos mais estruturados ou não. Dentro deste contexto, a seção a seguir aborda os três tipos conhecidos de processos: estruturado, semiestruturado e não estruturado.

2.1.2 Tipos de Processo

Em muitas organizações é possível encontrar processos com muitos pontos de decisões e com um amplo espectro de comportamentos permitidos. A aplicação da Mineração de Processos neste tipo de cenário geralmente resulta em modelos de

processos com elevado número de atividades e transições, sendo conhecidos também como processos não estruturados [1].

Três tipos diferentes de processos são contextualizados [6]: (1) estruturados ou “lasanha”, onde todas as atividades são repetitivas, controladas e possuem entradas e saídas bem-definidas; (2) semiestruturados, onde são conhecidos os requisitos de informações das atividades e é possível esboçar os procedimentos a serem seguidos, possuindo algumas características de autonomia e com decisões tomadas por julgamentos de seus executores; e (3) processos não estruturados ou “espaguete”, onde é difícil definir as pré-condições e as pós-condições de execução das atividades. Eles são conduzidos pela experiência dos executores e decisões autônomas sem regras ou procedimentos padronizados; eles são marcados pelo uso da intuição.

Um processo é considerado estruturado quando, com esforços limitados, é possível criar um modelo de processo em que mais de 80% dos eventos acontecem como planejado [15]. A Figura 3 permite observar que processos deste tipo possuem maior legibilidade de seus modelos. Logo, é possível analisa-los com maior facilidade dado a existência de um número reduzido de transições e atividades. Em contrapartida, essa visão é oposta quando comparado com a ilustração da Figura 4 que exibe um modelo não-estruturado.

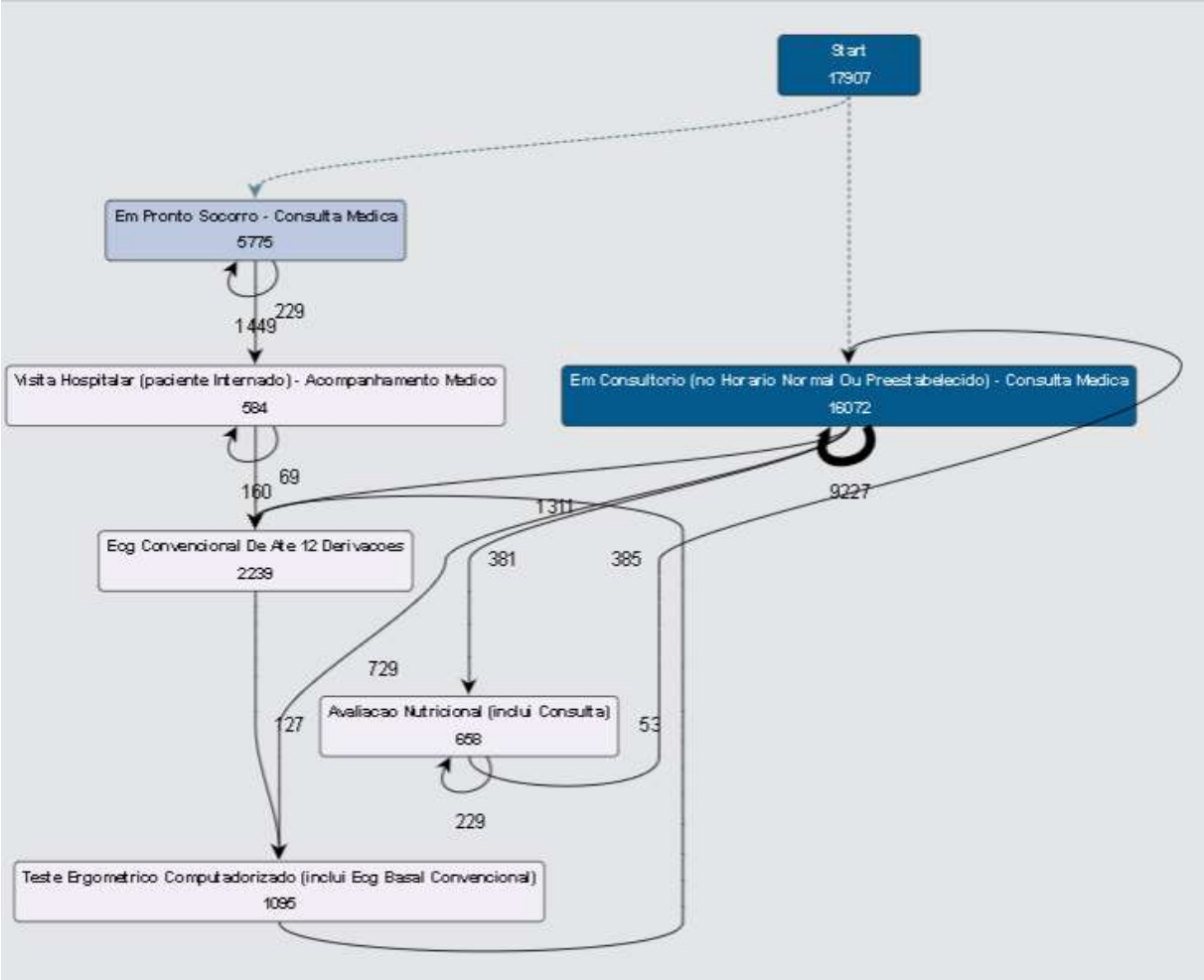


Figura 3 - Processo estruturado ou "Lasanha". Fonte: o Autor, 2017.

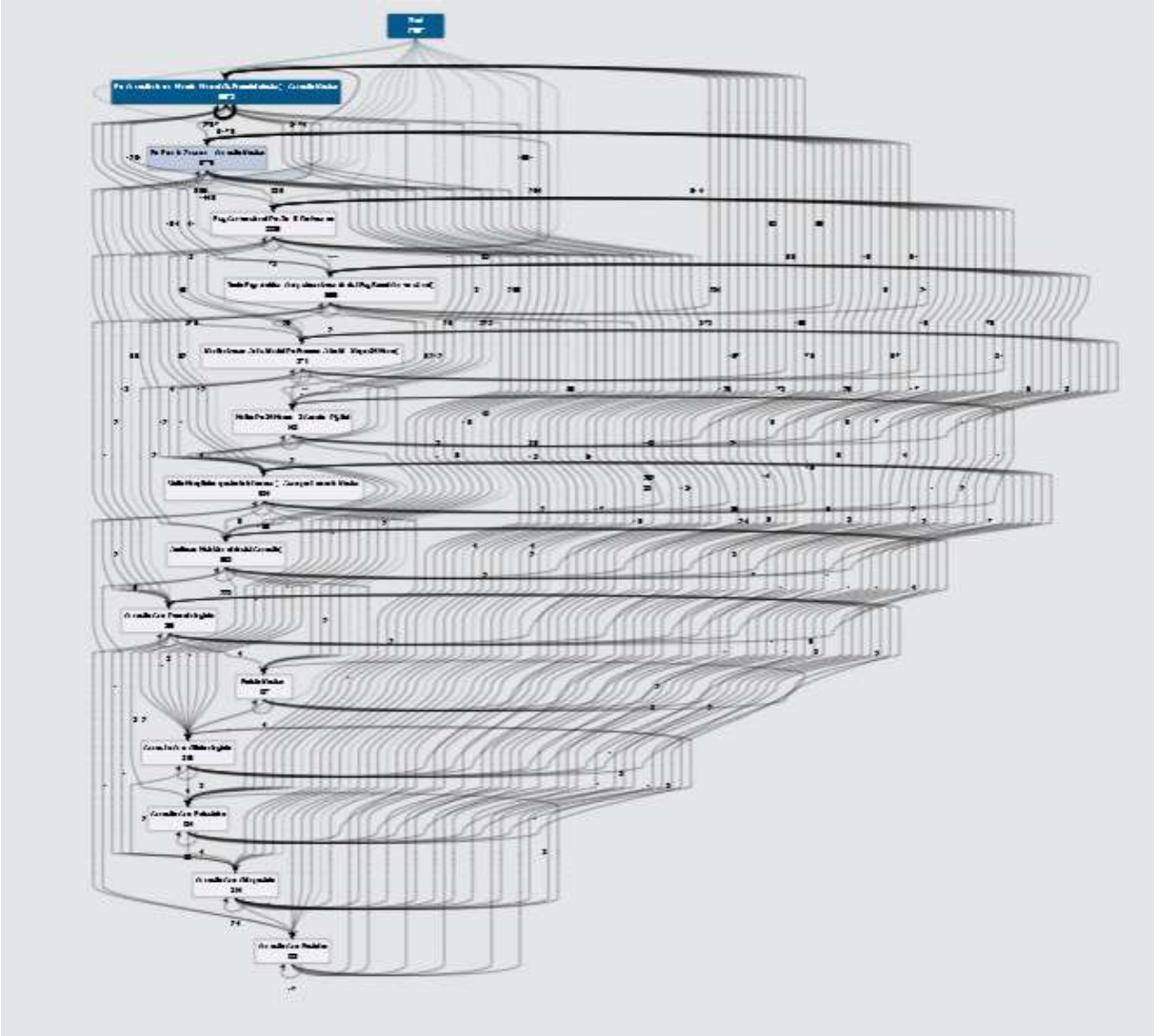


Figura 4 - Processo não estruturado ou "Espaguete'. Fonte: o Autor, 2017.

Ao contrário dos processos Lasanha, os processos Espaguete são mais difíceis de analisar. Porém, esses processos do ponto de vista da Mineração de Processos são muito interessantes, à medida que eles possuem um grande potencial de ganhos, face as diversas melhorias que podem ser realizadas vis-à-vis as padronizações, assim como as identificações de desvios e gargalos [15]. A aplicação de técnicas de agrupamento de instâncias de processos neste contexto diminui a complexidade de um modelo de processo não-estruturado, permitindo realizar uma análise individual dos modelos de processos, e consequentemente, melhorar a qualidade dos resultados obtidos da Mineração de Processos em ambientes flexíveis.

Outra representação para um processo é definida como modelo *flor* (cf. Figura 5), cuja representação permite qualquer ordem de execução de atividades em uma eventual reprodução do modelo de processo [6].

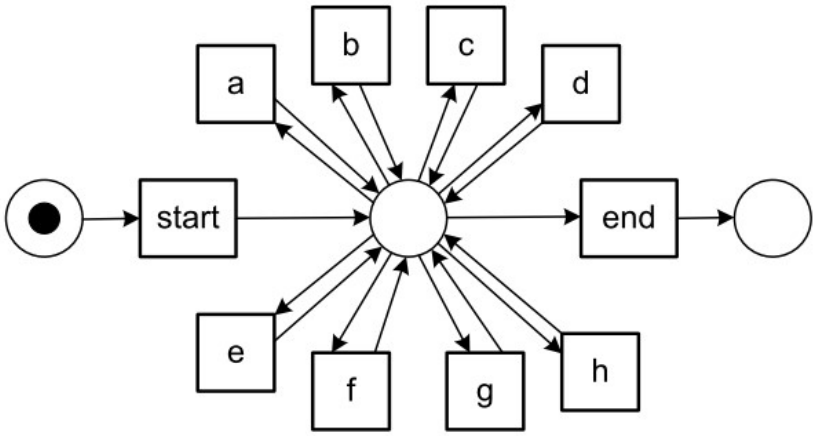


Figura 5 - Modelo flor. Fonte [6]

Depois desta breve contextualização sobre os tipos de processos, a próxima seção irá abordar tópicos relacionadas a Mineração de Processos, Técnicas de Mineração de Processos e Qualidade de Modelos.

2.2 Mineração de Processos

A Mineração de Processos, como já dito, é uma área de pesquisa relativamente nova e está relacionada a aprendizagem de máquina e mineração de dados. A ideia básica da Mineração de Processos é descobrir, monitorar e melhorar processos reais, extraindo conhecimentos de *logs* de eventos disponíveis em diversos sistemas de informação [6].

A Mineração de Processos, assim como as abordagens clássicas para a melhoria de processos, tais como: Melhoria Contínua de Processos (CPI), Gestão da Qualidade Total (TQM) e Seis Sigma possuem um objetivo em comum: “colocar um microscópio sobre os processos", com objetivo de identificar problemas com foco na otimização do desempenho operacional [6]. Seis Sigma, por exemplo, é uma metodologia abrangente fundamentada na estatística para identificar, isolar e eliminar

variações ou defeitos em processos [31]. Esta metodologia segue a abordagem DMAIC (sigla para os termos em inglês *Define, Measure, Analyse, Improve e Control*), que consiste em cinco etapas: (a) definir o problema a ser resolvido e as metas a serem alcançadas; (b) coletar os dados dos indicadores de desempenho; (c) analisar os dados para investigar e verificar as relações de causa e efeito; (d) melhorar o processo atual com base em análises retrospectivas; (e) controlar o processo para minimizar os desvios da meta [6]. A Mineração de Processos, neste contexto, pode ser aplicada em conjunto com os Seis Sigmas, especialmente nas etapas de análise e controle, a partir dos conhecimentos obtidos dos modelos de processos descobertos.

A Mineração de Processos também está relacionada a outra importante área de conhecimento, denominada como Gerenciamento de Processos de Negócio (BPM). BPM tem como objetivo apoiar os processos de negócios aplicando métodos, técnicas e suporte computacional para modelagem, controle, análise e melhoria de processos operacionais realizados por atores humanos, atores computacionais, documentos e outras fontes de informação [32]. Dentro do BPM, a Análise de Processos de Negócio (BPA) é uma subárea que inclui desenhos de modelos de negócios, técnicas de simulação, diagnósticos, verificações de desempenho e análises de processos [33]. Outra subárea concerne ao Monitoramento de Atividades de Negócio (BAM), cujo objetivo é o diagnóstico de desempenho dos processos em sistemas de BPM, utilizando indicadores e regras de controle. BAM tem objetivos semelhantes aos da Mineração de Processos, especificamente na análise de aprimoramento [32].

A Mineração de Processos encerra um conjunto de abordagens, cujo objetivo principal é extrair conhecimentos de *logs* de eventos armazenados em diversos sistemas de informações, onde é possível: 1) descobrir modelos que representam o processo analisado; 2) verificar a conformidade dos eventos em relação ao modelo descoberto; 3) ampliar e expandir o modelo descoberto combinando informações de gargalos, desempenho, recursos [6], etc.

Por fim, no restante desta seção serão descritos conceitos relacionados ao *log* de eventos, Mineração de Processos, técnicas de mineração, tipos de processos analisados, e métricas de avaliação de modelos de processos.

2.2.1 *Log* de Eventos

Log de eventos representam instâncias de processos armazenados em sistemas de informações e podem ser vistos como um conjunto de traços, onde cada traço descreve o ciclo de vida de um caso em particular, que por sua vez é composto por eventos [7]. A Tabela 1 ilustra o fragmento de um *log* de eventos, onde cada linha corresponde a um evento. Os eventos se referem a dois casos (205061, 205085) e possuem atributos que representam a atividade que foi executada, quando ela ocorreu e o caso que a mesma está relacionada.

Tabela 1 - *Log* de eventos. Fonte: o Autor, 2017.

Caso	Data	Atividade	Recurso	Escritório
205061	03/01/2016 15:40:00	Criar ordem de venda	Eduardo	Vendas
205061	03/01/2016 15:55:20	Adicionar Produtos	Maria	Vendas
205061	03/01/2016 15:55:25	Adicionar Produtos	Maria	Vendas
205085	03/01/2016 15:42:00	Criar ordem de venda	Eduardo	Vendas
205085	03/01/2016 15:45:19	Adicionar Produtos	Maria	Vendas
205085	03/01/2016 15:49:35	Faturar pedido	Maria	Faturamento
205061	03/01/2016 16:01:46	Solicitar Orçamento Técnico	João	Vendas
205061	03/01/2016 16:09:01	Enviar Orçamento	João	Engenharia
...	...	...	...	...

No *log* de eventos da Tabela 1 é possível verificar que existem atributos adicionais, como Recurso e Escritório. Estes atributos permitem analisar os processos em diferentes perspectivas, respondendo diferentes questões no momento da análise. Por exemplo, é possível verificar que a atividade do caso “205061” foi realizada na Data “03/01/2016 15:40:00” pelo recurso “Eduardo”. Outros comportamentos também podem ser observados, por exemplo, a atividade “Adicionar Produtos” sempre é precedida da atividade de “Criar ordem de vendas” e ambas são executadas no

escritório de “Vendas”. A partir dos *logs* de eventos pode-se conduzir quatro tipos de Mineração de Processos que serão descritos na próxima subseção.

2.2.2 Tipos de Mineração de Processos

A partir de um conjunto de traços presente nos *logs* de eventos pode-se conduzir quatro tipos de Mineração de Processos [6]: descoberta, conformidade, aprimoramento e suporte operacional (cf. Figura 6).

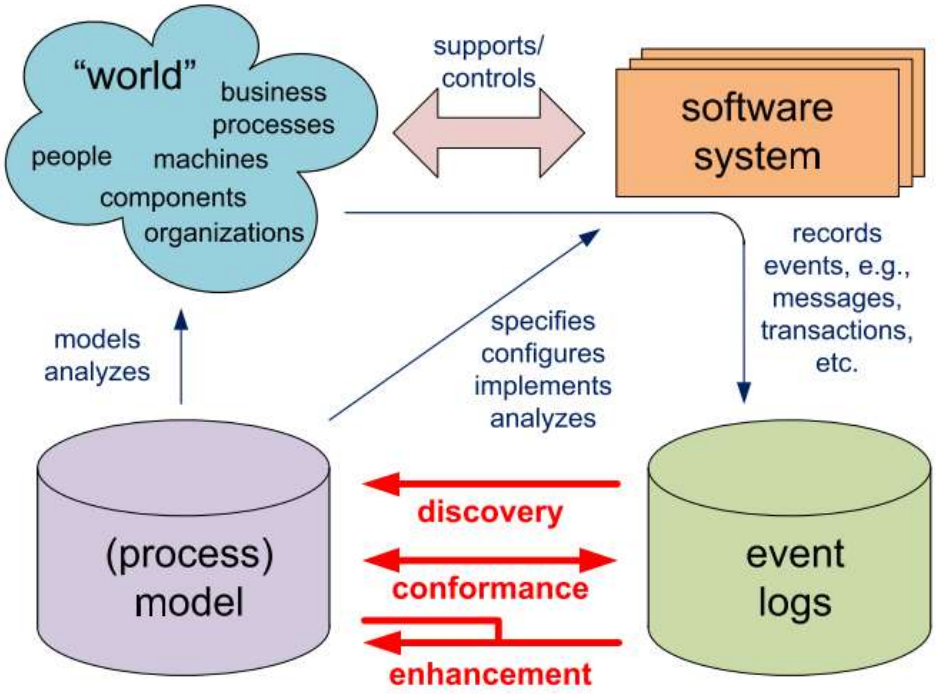


Figura 6 - Tipos de mineração de processos. Fonte: [6].

O processo de descoberta é responsável por produzir modelos de processos sem utilizar qualquer tipo de informação prévia do processo [6]. Ao longo dos anos diversas técnicas surgiram com este propósito, sendo o *Alpha Algorithm* uma abordagem precursora. O resultado obtido da aplicação deste algoritmo, por exemplo, é um modelo de processo baseado em uma Rede de Petri (cf. Figura 7).

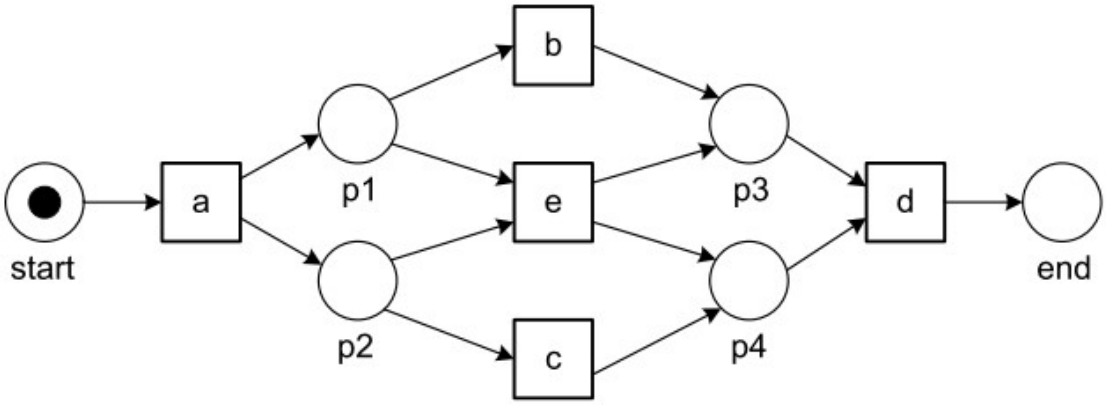


Figura 7 – Modelo de processo descoberto. Fonte: [6].

Após a descoberta de um processo, a análise de conformidade tem como objetivo principal comparar um modelo de processo existente com o modelo descoberto a partir de um *log* de eventos. Tal iniciativa pode ser utilizada para verificar se o processo real, registrado em *log*, está em conformidade com o modelo previamente estabelecido e vice-versa. Desta forma, pode-se detectar, localizar e explicar desvios em um processo, bem como mensurar a importância e gravidade dos mesmos [6].

O terceiro tipo de Mineração de Processos é chamado de aprimoramento ou extensão e visa melhorar e enriquecer um modelo existente de processo a partir de informações do processo que podem ser extraídas do *log* de eventos. Um exemplo de aprimoramento é a combinação de modelos de processos com dados de desempenho, obtidas por exemplo na data de início e término das atividades [7]. A Figura 8 ilustra a aplicação da extensão em um modelo de processo, onde é possível observar algumas atividades e transições com uma maior ênfase. Nesse caso em específico, as atividades nas cores acentuadas para o vermelho representam tempo de execução maior em relação as outras atividades, caracterizando possíveis gargalos no processo.

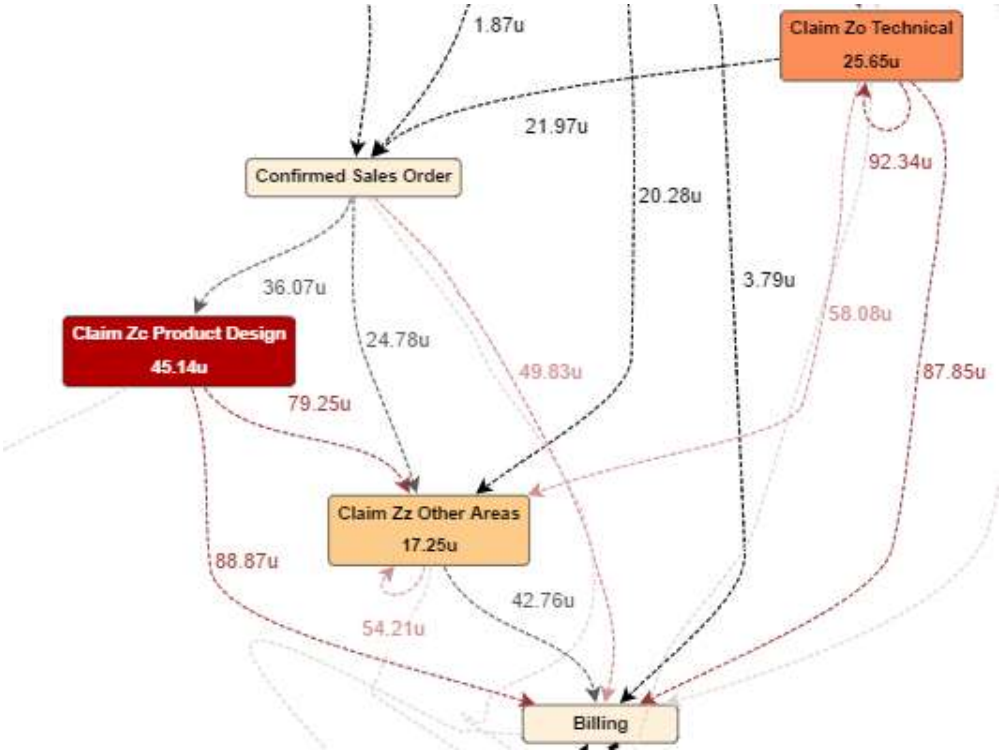


Figura 8 - Modelo de processo estendido. Fonte: o Autor, 2017.

Por fim, o último tipo de Mineração de Processos é chamado de suporte operacional, cuja principal diferença em relação aos tipos de Mineração de Processos descritos anteriormente está relacionada a análise realizada de forma *online*. Desta forma, com base em modelos de processos descobertos pode-se verificar, prever, ou recomendar atividades em casos que estão ocorrendo em uma configuração *online*, influenciando as execuções dos processos em tempo real [7]. Com base nos tipos de Mineração de Processos, serão apresentadas a seguir algumas técnicas que permitem colocar em prática a operacionalização de diferentes tipos de mineração.

2.2.3 Técnicas de Mineração de Processos

As técnicas de descoberta de processos incluem vários algoritmos de mineração, como por exemplo: *Alpha Algorithm* [17*], Heuristic Miner* [19] [21*], Fuzzy Mine*r [22], *Genetic Miner* [23], *Integer Linear Programming Miner* [24], e *Mineração Declarativa* [25]. A primeira técnica considera a causalidade dos traços no *log* de eventos de um

fluxo de trabalho [17]. Ela descobre o modelo de uma instância de processo executada e contida nos *logs* de eventos, representando-o por meio de uma *Rede de Petri*.

O *Heuristic Miner* estende a ideia do *Alpha Algorithm*. Ele é menos sensível a *log* de eventos incompletos e com presença de ruídos. Essa robustez decorre do uso de filtros que removem as atividades e transições menos frequentes; tais filtros são baseados na frequência dos traços. A tarefa de mineração é dividida em três passos [21]: (1) construir uma tabela de dependências com base em métricas de frequência; (2) construir um grafo de dependência a partir da tabela de dependência, onde as divisões e junções de processos são classificadas utilizando uma representação de matriz causal; e por fim, (3) construir um modelo de processo utilizando a tabela e o grafo de dependências, representando-o na forma de uma *Rede Causal*. Esta técnica gera modelos de processos com qualidade mesmo quando há ruídos nos *logs* [30].

A abordagem *Fuzzy Miner* [22], permite analisar o modelo descoberto, dando ênfase as variantes mais relevantes de acordo com medidas de significância. Quando um processo não estruturado é analisado, filtros com base nestas medidas são úteis para aumentar a compreensão do processo. Esta abordagem se baseia em quatro pilares: (1) Agregação: o número de elementos visualizados é limitado; (2) Abstração: as informações insignificantes no contexto escolhido são omitidas; (3) Ênfase: informações mais significantes são destacadas, i.e., os caminhos com maior tempo de espera ou frequência; (4) Customização: diferentes visões podem ser geradas de acordo com as perspectivas de análise. A visualização do processo é dada por um diagrama de transições, não representando paralelismo, junções e divisões no processo.

Por fim, uma das mais recentes e promissoras técnicas de descoberta de modelos de processos é conhecida como *Inductive Miner.* Ela é capaz de garantir que os modelos descobertos são *fit* (i.e., consegue reproduzir todo comportamento observado) e *sound* (i.e. livre de *deadlocks* e outras anomalias) [58]. Esta abordagem representa o modelo de processo descoberto como um conjunto de blocos estruturados utilizando a representação de árvore, convertível em uma *Rede de Petri*. A técnica de dividir e conquistar é utilizada para decompor o problema de descoberta para acelerar o processo, onde o *log* de eventos L é dividido em n sub*logs* obtidos

pela divisão de L de modo a descobrir n sub-processos. A ideia básica das divisões (cf. Figura 9) é encontrar estruturas que indicam operadores dominantes que caracterizam o comportamento de uma sequência de atividades. Estes operadores representam pontos de decisões do processo como *E, OU, paralelismo e loops*. Por exemplo, se um grupo de atividades é precedido por outro grupo de atividades, mas nunca ao contrário, pode ser inferido que esses grupos estão em sequência, e consequentemente, o *log* de eventos é decomposto em dois grupos de atividades [58]. Desta forma, os *sublogs* são decompostos até eles se referirem a uma única atividade.

Algorithm 1 Inductive miner

```
1: procedure RECURSIVELY(a)
2:     Find most significant split in event log
3:     Detect operator
4:     Continue on sublogs
5: end procedure
```

Figura 9 - Algoritmo *Inductive Miner*. Fonte: [58]

O *Inductive Miner* possui uma família de técnicas baseadas nos conceitos apresentadas anteriormente. Em tal família existem variações da técnica e dos operadores elaborados para lidar com comportamentos infrequentes, ruídos e grandes *logs* de eventos. Esta técnica é considerada atualmente o estado da arte referente as técnicas de descoberta de modelos de processos devido sua flexibilidade, qualidade dos modelos gerados e escalabilidade [58].

2.2.4 Ferramentas

Diversas ferramentas de Mineração de Processos foram construídas nos últimos anos com diferentes objetivos de aplicação. No meio acadêmico, destaca-se a plataforma de Mineração de Processos chamada ProM[1]. Ela implementa diversas técnicas de Mineração de Processos [6] e suporta *log* de eventos nos formatos XES, MXML e CSV. Pode-se usar tais formatos em conjunto com técnicas*: Alpha Algorithm* [17], *Heuristic Miner* [19] [21], *Fuzzy Miner* [22], *Genetic Miner* [23], *Integer Linear Programming Miner* [24], *Mineração Declarativa* [25] e *Inductive Miner* [58] para

[1] Framework de Mineração de Processos. Disponível em: http://www.promtools.org/

descobrir modelos e extensões de modelos. Medidas para avaliação de conformidade como *Fitness* e *Behavioral Appropriateness* [14] também são disponibilizadas, assim como outras abordagens de *streaming* de eventos. A Figura 10 ilustra a aplicação de uma técnica de descoberta na ferramenta ProM.

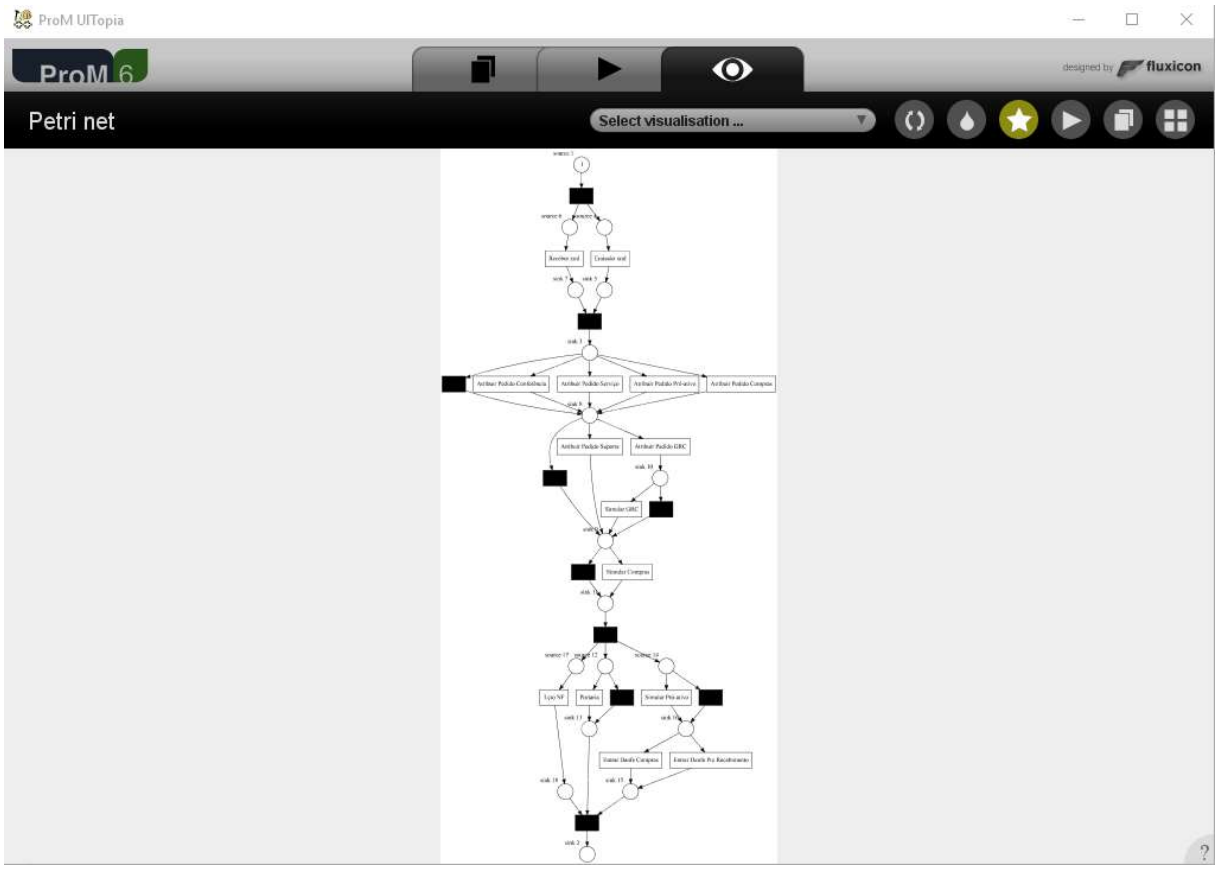


Figura 10 - Visualização de um modelo de processo descoberto na ferramenta ProM.

2.3 Métricas de Qualidade

Determinar a qualidade de um modelo descoberto a partir de técnicas de mineração de processos é uma tarefa difícil e ela se caracteriza por muitas dimensões. Estas dimensões definem o quanto um modelo descoberto consegue refletir de forma adequada o comportamento existente no *log* de eventos. A Figura 11 ilustra quatro dimensões de qualidade apresentadas [6]: *fitness*, simplicidade, generalização e precisão.

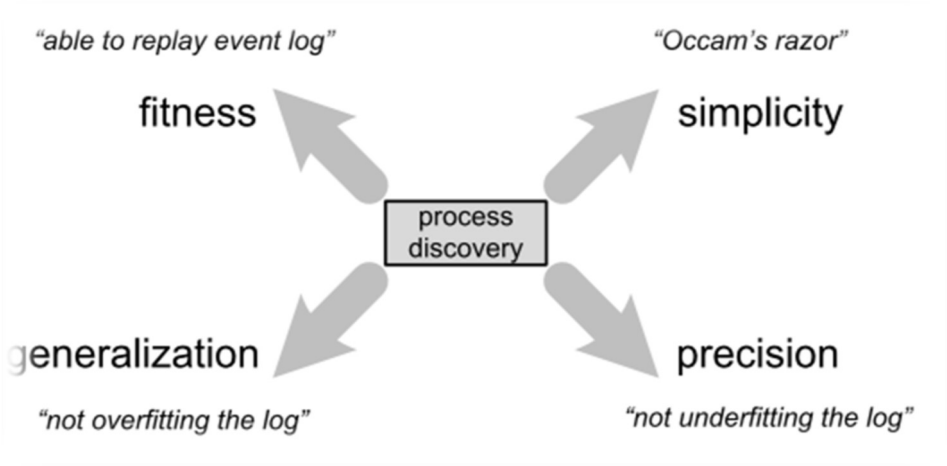


Figura 11 – Equilíbrio das quatro dimensões. Fonte: [6].

As técnicas de Mineração de Processos buscam encontrar um equilíbrio entre as quatro dimensões para gerar modelos de processos descobertos com qualidade [6], a saber: (1) *Fitness* se refere aos modelos que conseguem refletir o comportamento observado no *log* de eventos, de modo que, um modelo é caracterizado como um perfeito *fitness* se todos os traços de um *log* puderem ser reproduzidos do começo ao fim; (2) *Simplicidade* se refere ao princípio de *Occam's Razor*, que afirma que "não se deve aumentar, além do necessário, o número de entidades necessárias para explicar algum fenômeno". Isto significa que o melhor modelo é o mais simples que consegue explicar o comportamento de um *log*. (3) *Precisão* estabelece que um modelo preciso é aquele que não permite muitos comportamentos, caso contrário ele é caracterizado como *underfitting,* i.e., o modelo é muito generalista permitindo comportamentos muito diferentes daqueles existentes no *log*. (4) *Generalização* define que um modelo não deve restringir-se a comportamentos muito específicos observados no *log*; estes que por sua vez são denominados como *overfitting*.

Neste contexto, as abordagens propostas na literatura visam mensurar a qualidade de modelos de processos baseado nas dimensões de qualidade. As próximas subseções irão apresentar medidas comumente utilizadas nas validações de técnicas de Mineração de Processos.

2.3.1 Fitness

A abordagem precursora para mensurar o *fitness* de um modelo de processo descoberto foi apresentada em [14] e consiste em tentar reproduzir as sequências de eventos do *log* no modelo de processo descoberto; técnica também chamada de *replay*. Em cada execução de uma sequência de eventos no modelo de processo são contabilizados *tokens*, a saber: (1) *tokens* que foram criados artificialmente no modelo devido transições existentes não terem sido habilitadas; (2) *tokens* faltantes, i.e. que foram produzidos na reprodução do log de eventos no modelo, mas não foram consumidos, indicando que o processo não foi devidamente concluído (3) *tokens que foram* produzidos na reprodução do *log* de eventos no modelo; (4) *tokens* que foram consumidos ao longo da reprodução do *log* de eventos no modelo.

A Figura 12 ilustra o processo de contabilização dos *tokens* para calcular o *fitness*, na qual duas sequências de eventos são reproduzidas no modelo de processo. Na primeira sequência de eventos são contabilizados m = 0, r = 0, c = 7 e p = 7, representando que foi possível reproduzir toda a sequência de eventos no modelo sem *tokens* faltantes—m = 0—ou restantes—r = 0. Na segunda sequência de eventos foram contabilizados m = 1 referindo-se ao *token* faltante c7 entre a transição das atividades G para H.

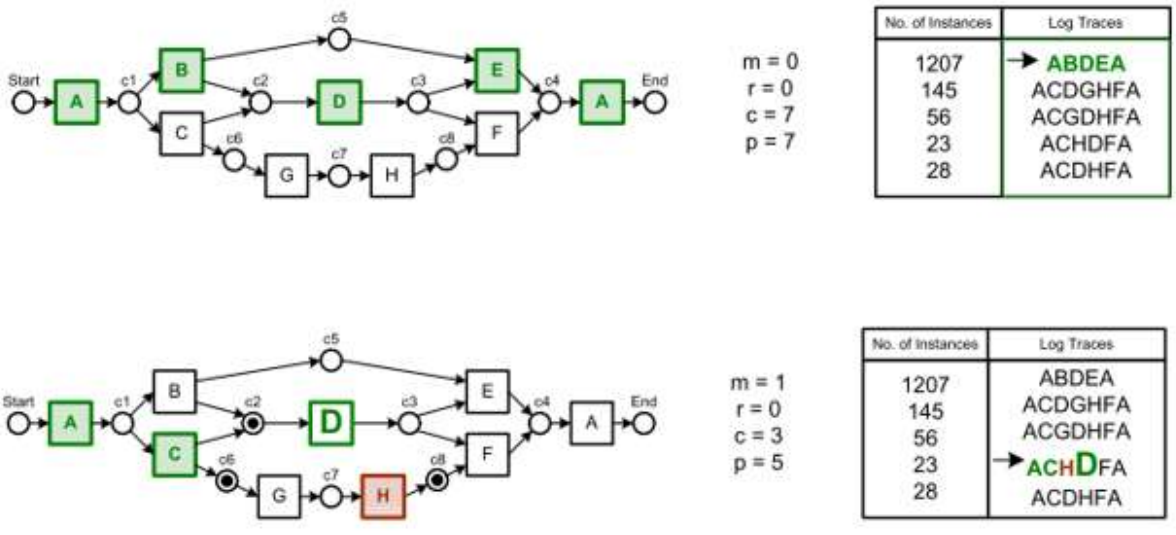


Figura 12 - Replay do *log* de eventos no modelo, onde *m, r, c, p* denotam respectivamente *tokens* ausentes, restantes, consumidos e produzidos Fonte: [14].

Contabilizados o número de *tokens* ausentes, restantes, consumidos e produzidos, a métrica f proposta por [14] é definida pela equação E1.

$$f = \frac{1}{2}\left(1 - \frac{\sum_{i=1}^{k} n_i m_i}{\sum_{i=1}^{k} n_i c_i}\right) + \frac{1}{2}\left(1 - \frac{\sum_{i=1}^{k} n_i r_i}{\sum_{i=1}^{k} n_i p_i}\right) \quad \text{E1}$$

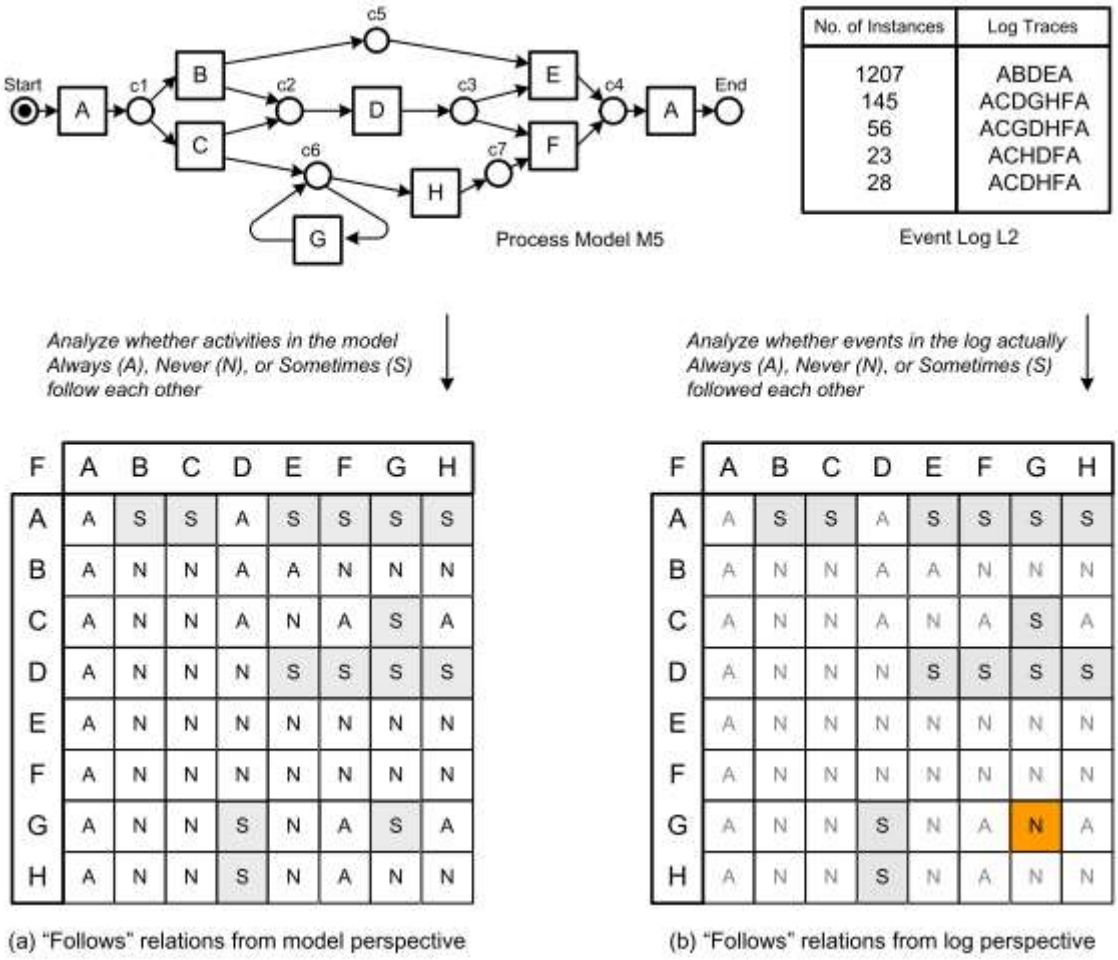
Onde k é o número de diferentes traços. Para cada traço i $(1 \leq i \leq k)$, n_i é o número de instâncias de processo combinadas com o traço corrente, m_i é o número de *tokens* ausentes, r_i representa o número de *tokens* restantes, c_i o número de *tokens* consumidos, por fim p_i é o número de *tokens* produzidos durante a repetição dos *logs* no traço [14]. Quanto mais próximo de 1 for o valor de *f* melhor é a qualidade do modelo de processo.

A medida de *Fitness* deve levar em consideração a severidade do desvio entre as instâncias de processos existentes no *log* de eventos e no modelo de processo descoberto, onde um desvio pode se manifestar em atividades omitidas ou em atividades inseridas que não aparecem no modelo, e que podem acontecer durante a execução do processo [38]. Neste contexto ressalta-se que técnicas clássicas [14] como apresentada anteriormente, utilizam determinadas heurísticas que em alguns casos apresentam estimativas incorretas de *Fitness*. Propôs-se aqui uma abordagem que permite calcular o *Fitness* lidando com atividades não observáveis, ignoradas e inseridas, levando em conta a sua importância. De maneira simplista, tal abordagem transforma um traço (i.e., registros de eventos contidos em uma instância de processo) em uma Rede de Petri e reduz o problema do cálculo de conformidade/*Fitness* por meio da interseção do traço com o modelo de processo descoberto, ambos representados por uma Rede de Petri. No modelo de processo descoberto, atributos de custo são adicionados para determinar o preço de ignorar atividades e o custo por tipo de atividade para atividades inseridas. Com base nos custos calculados e interseção dos modelos, a medida de *Fitness* é calculada.

2.3.2 Precisão

O método *Behavioral Appropriateness* visa medir a *Precisão* de modelos de processos [14]. Ele avalia os comportamentos permitidos nos modelos de processos que não se encontram na execução*/replay* dos *logs* de eventos. A motivação para esta métrica está relacionada aos modelos de processos muito generalistas, que permitem sequências de execuções que não condizem com a realidade. Ainda, a partir do conjunto de sequências de atividades pode-se derivar relações comparáveis denominadas de “Sequente” e “Precedente”, tanto em um modelo de processo, quanto em um *log* de eventos. Isso é feito de tal modo que, observado um conjunto de sequências, pode-se também determinar se duas atividades, *Sempre*, *Nunca* ou *Algumas vezes* seguem ou precedem umas às outras. As relações *Sempre* e *Nunca* descrevem relações fortes, e as relações *Algumas vezes* capturam as variabilidades no comportamento.

Desta forma, a ideia da métrica é comparar as variabilidades do comportamento permitido pelo modelo e o comportamento observado no *log* com base nas relações “algumas vezes seguem”—denominadas como S_F—e “algumas vezes precedem”—denominadas S_P. A Figura 13 apresenta as relações em duas perspectivas: 1) modelo, onde são analisadas as possíveis sequências de execução; e 2) *log* de eventos, onde é possível analisar as sequências de execuções observadas.



F	A	B	C	D	E	F	G	H
A	A	S	S	A	S	S	S	S
B	A	N	N	A	A	N	N	N
C	A	N	N	A	N	A	S	A
D	A	N	N	N	S	S	S	S
E	A	N	N	N	N	N	N	N
F	A	N	N	N	N	N	N	N
G	A	N	N	S	N	A	S	A
H	A	N	N	S	N	A	N	N

F	A	B	C	D	E	F	G	H
A	A	S	S	A	S	S	S	S
B	A	N	N	A	A	N	N	N
C	A	N	N	A	N	A	S	A
D	A	N	N	N	S	S	S	S
E	A	N	N	N	N	N	N	N
F	A	N	N	N	N	N	N	N
G	A	N	N	S	N	A	N	A
H	A	N	N	S	N	A	N	N

Figura 13 - Relações derivadas do modelo de processo M5 e do *log* de eventos L2. Fonte: [14].

Neste contexto, tendo S_F^m uma relação S_F e S_P^m uma relação S_P para modelo de processo, e S_F^l uma relação S_F e S_P^l uma relação S_P para o *log* de eventos a métrica *Advanced Behavioral Appropriateness* [14] é definida pela equação E2.

$$ba = \left(\frac{|S_F^l \cap S_F^m|}{2 \,.\, |S_F^m|} + \frac{|S_P^l \cap S_P^m|}{2 \,.\, |S_P^m|} \right) \quad \text{E2}$$

A abordagem *Advanced Behavioral Appropriateness* demanda um elevado custo computacional [30]. Isto decorre das exaustivas simulações para analisar um modelo de processo; sua aplicação é inviável para grandes *logs* de eventos. Baseado na dificuldade das abordagens existentes em calcular a precisão de modelos complexos e com grandes *logs* de eventos, uma proposta alternativa apresentada para medir *Precision* de maneira escalável com capacidade de lidar com modelos de processos complexos e com grandes *logs* de eventos, utilizando princípios da abordagem de dividir e conquistar [34]. Ao invés de comparar o comportamento completo envolvendo todas as atividades, o problema é decomposto comparando o comportamento de subconjuntos de atividades. Desta forma, para cada subconjunto, é calculada *Precision* e a média da *Precision* de todos os subconjuntos resulta no valor final da medida.

2.4 Agrupamento em Mineração de Processo

Um agrupamento é caracterizado como uma classificação não supervisionada em mineração de dados. Tal abordagem tem como principal objetivo agrupar conjuntos de instâncias de processos em grupos significativos [37]. O processo de agrupamento é dividido em três grandes etapas (cf. Figura 14): 1) seleção e extração de características; 2) definição da similaridade entre instâncias; e 3) agrupamento de instâncias.

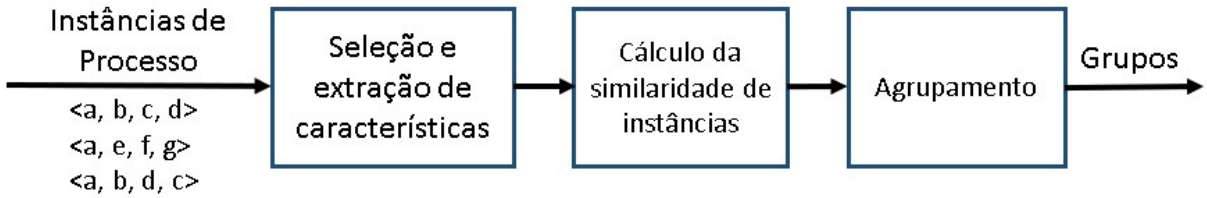


Figura 14 - Estágios do processo de agrupamento. Fonte: [37].

A etapa *de seleção e extração de características* é responsável por identificar as características mais eficientes a serem utilizadas na tarefa de agrupamento e extrai-las utilizando um ou mais métodos de transformações para produzir características mais relevantes. A etapa *cálculo da similaridade de instâncias* atribui

um grau de similaridade entre os padrões/instâncias; isso é feito por meio da aplicação de métricas faces as características dos dados. Por fim, a última etapa se refere ao agrupamento dos padrões similares em grupos de acordo com o grau de similaridade definido na etapa anterior [37].

Algoritmos tradicionais de Mineração de Processo não lidam adequadamente com processos não estruturados. Eles geram modelos de processos de difícil compreensão [29]. Dentro deste contexto, o agrupamento é utilizado como etapa de preparação para a descoberta de modelos de processos. Tal preparação visa melhorar a acurácia e compreensão dos modelos ([1] [29]), descobrir variantes e desvios do processo analisado [5], de modo que as instâncias similares são agrupadas no mesmo conjunto, gerando modelos de processos mais estruturados. Esta ideia é ilustrada na Figura 15, onde observa-se que modelos de processos resultantes de cada subconjunto/agrupamento é visualmente mais compreensível que os modelos de processos descobertos a partir do conjunto completo.

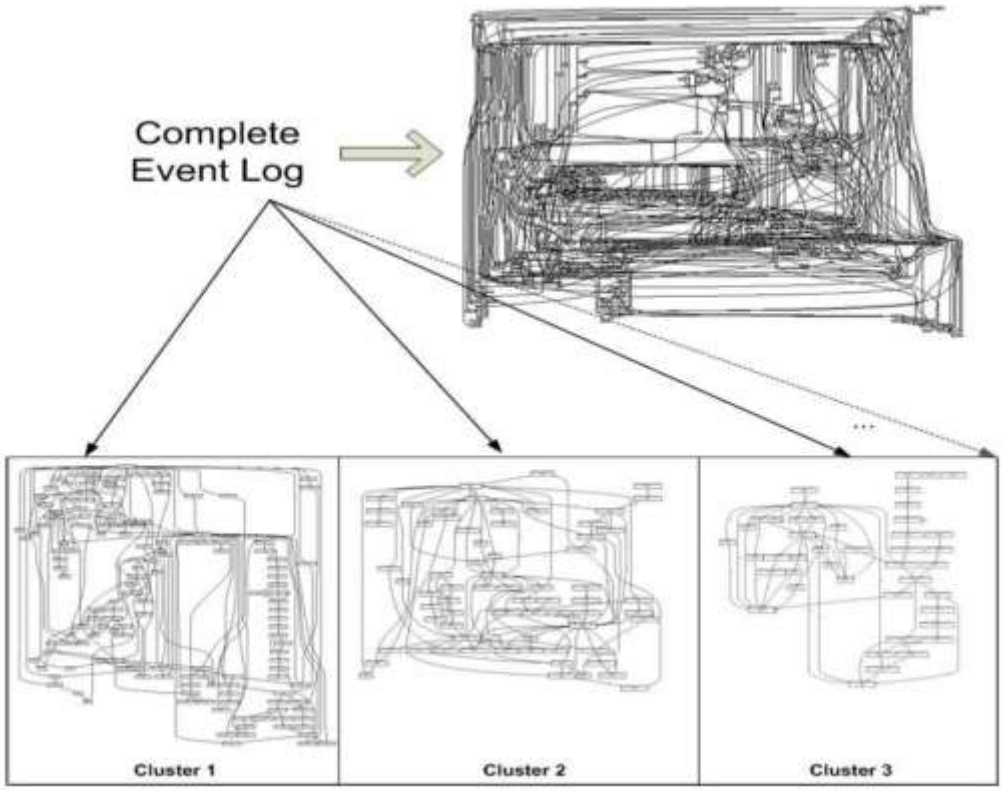


Figura 15 - Agrupamento de instâncias de processos. Fonte: [29].

Com o objetivo de melhorar os resultados da Mineração de Processos, [29] apresentou uma abordagem de agrupamento de traços baseada no agrupamento hierárquico, na qual inicia cada dado (neste caso, um traço de um evento) em diferentes grupos e iterativamente combina-os até obter um único grupo. Para a avaliação da qualidade do agrupamento foram utilizadas medidas de *fitness* e de *simplicidade* sob a ótica da qualidade de modelos de processos.

Outra abordagem de agrupamento de traços foi proposta em [1], baseada nos perfis dos dados contidos no *log* de eventos, ou seja, conjuntos de características que representam diferentes perspectivas de análise presentes em um *log* de eventos. Estes perfis são definidos pelos nomes das atividades, transições, metadados de casos e atributos de eventos. Uma matriz com os perfis é construída para possibilitar o cálculo da similaridade entre as sequências de eventos. Neste caso foi utilizada para realizar o agrupamento dos traços uma combinação de técnicas, envolvendo, de um lado, o cálculo de distância Euclidiana e, de outro, o algoritmo de agrupamento *Self-Organizing Map* (SOM). A qualidade do agrupamento foi avaliada por meio das medidas de qualidade *fitness* e *behavioral appropriateness* [14].

Com o objetivo de descobrir variantes e desvios do processo, [5] apresentou uma abordagem de agrupamento de traços baseada em um algoritmo de agrupamento *markoviano*. Esta abordagem utiliza matrizes estocásticas para representar as probabilidades de transições entre as atividades dos traços de eventos. Neste caso, a entrada do algoritmo é uma matriz de similaridade de traços que contem a similaridade entre todas as instâncias de processos presentes em um *log* de eventos. A saída são grupos de traços constituídos conforme a similaridade entre instâncias. A avaliação de tal abordagem foi realizada por meio da métrica *fitness.* Adicionalmente, os traços foram categorizados em diferentes cenários conforme as características dos atributos presentes no *log* de eventos; isto foi feito para permitir avaliar a entropia e taxa de divisão. No caso da entropia, quanto mais cenários estiverem agrupados maior será o valor. Na taxa de divisão foram avaliados em quantos agrupamentos diferentes foram assinalados um mesmo cenário. Idealmente, nenhum agrupamento deve conter mais de um cenário e nenhum cenário deve ser dividido em mais de um agrupamento, ou seja, a entropia é 0 e a taxa de divisão é 1.

Como pôde ser observado, a eficiência do agrupamento quando relacionada a Mineração de Processo é avaliada sob perspectivas das métricas de qualidade de modelos de processos, de tal forma que a eficiência das abordagens é determinada pelo equilíbrio entre as dimensões de qualidade [35].

2.4.1 Técnicas de Agrupamento

O agrupamento de instâncias pode ser distinguido em vários tipos e de acordo com as técnicas empregadas para a construção dos agrupamentos. Tais técnicas podem ser classificadas como *hierárquica* versus *particional*, *exclusiva* versus *sobreposta* versus *fuzzy*, e *completa* versus *parcial* [59]. Essas são contextualizados como segue:

- **Hierárquica versus *Particional*:** o agrupamento hierárquico consiste na divisão do conjunto de instâncias em subconjuntos não sobrepostos, i.e., cada instância pertence apenas a um único subconjunto. Caso o agrupamento permita que conjuntos tenha subconjuntos, é obtido então o agrupamento hierárquico, onde os subconjuntos são aninhados e organizados como uma árvore.
- **Exclusivo versus Sobreposição versus *Fuzzy*:** O agrupamento exclusivo define que uma instância deve ser atribuída a um único grupo, ao contrário do agrupamento com sobreposição, onde as instâncias podem simultaneamente estar em mais de um grupo. Já o agrupamento *fuzzy* define que as instâncias são atribuídas a todos os grupos, de forma que a definição da importância de cada instância para cada grupo é dada por um peso entre 0 e 1.
- **Completo versus Parcial:** O agrupamento completo garante que toda instância é assinada a um grupo, ao contrário do agrupamento parcial. A motivação para o agrupamento parcial é que algumas instâncias podem não pertencer a grupos bem-definidos.

Uma das técnicas empregadas para realizar o agrupamento de grandes quantidades de instâncias é assinalar as instâncias ao grupo pertinente sem recalcular

a centroide [37]. Esta técnica se baseia na premissa que quando uma instância é assinalada a um grupo ela não afeta os conjuntos significativamente, de modo que não é necessário armazenar todas as instâncias em memória. O custo computacional desta abordagem é $O(n)$, onde n representa o número de instâncias a serem agrupadas. A decisão de atribuir uma instância *A* a um dado grupo *B* é feita com base na similaridade entre *A* e *B*. A ideia básica do algoritmo é ilustrada na Figura 16.

Algoritmo de agrupamento incremental

1. Atribuir a primeira instância a um conjunto;
2. Para cada nova instância;
   - **2.1.** Atribuir a um dos conjuntos existentes ou a um novo conjunto de acordo com o limiar estabelecido da distância da nova instância com os centroides dos conjuntos existentes;
3. Repetir o passo 2 até que todos as instâncias sejam agrupadas

Figura 16 - Algoritmo de agrupamento incremental. Fonte [37]

2.5 Considerações sobre o Capítulo

Neste capitulo foram estabelecidos os conceitos necessários para planejamento e execução da pesquisa. Foram apresentados diferentes tipos de processos que podem ser analisados por meio da abordagem de Mineração de Processo e métricas para avaliação da qualidade do resultado da mineração de processo.

Por último, a aplicação de técnicas de agrupamento de instâncias de processo em conjunto com a Mineração de Processo foi contextualizada com o objetivo de melhorar a compreensão dos modelos de processos gerados.

CAPÍTULO 3 - MEDIDAS DE SIMILARIDADE

A similaridade não é definida diretamente por uma fórmula. Ela é dada por um conjunto de pressupostos. Tais pressupostos são divididos em três proposições [16]: (1) a similaridade entre A e B está relacionado com a sua semelhança, de tal forma que, quanto mais elementos em comum, mais similares eles são; (2) a similaridade entre A e B está relacionada com as diferenças entre eles, onde quanto mais diferenças eles têm, menos similares eles são; e (3) a máxima similaridade entre A e B é atingida quando ambos são idênticos.

Diversas medidas de similaridade de modelos de processo foram propostas na literatura com diferentes objetivos, tais como: gerenciamento de repositórios de modelos, descoberta de serviços, conformidade de processos, simplificação de mudanças, reuso e entre outros. Estas técnicas podem ser segmentadas em 4 grupos de acordo com as técnicas empregadas [3]: (1) medidas baseadas na correspondência entre nós e arestas; (2) medidas baseadas na distância de edição de grafos; (3) medidas baseadas na análise das dependências causais entre as atividades; e (4) medidas baseadas na comparação do conjunto de traços ou *logs* de eventos.

A seguir serão contextualizadas as propriedades de medidas de similaridade derivadas das proposições citadas acima e em seguida, as técnicas de medidas de similaridades de modelos de processos.

3.1.1 Preliminares

Nesta subseção serão apresentadas algumas notações importantes e conceitos utilizados neste estudo antes da contextualização das medidas de similaridade de modelos de processos.

Um modelo de processo definido como M_0 pode ser comparado com outro modelo M_1. Cada modelo é descrito como um grafo dirigido (N, E) que contém um

conjunto N de nós e um conjunto E de arestas, onde $M_0 = (N, E)$. Nós são abstrações para transições e lugares um uma rede de Petri e arestas para os arcos. Atividades A são um subconjunto de nós $A \subseteq N$ que possuem rótulos textuais atribuídos.

3.1.2 Propriedades de medidas de distância e similaridade

Uma revisão sistemática sobre as técnicas utilizadas para definir e calcular a similaridade entre modelos de processos foi feita por [3]. Ela visava compreender como diferentes técnicas medem a similaridade dentro de um mesmo conjunto de modelos de processos. Para avaliar tais técnicas foram estabelecidas propriedades de medidas de distância e similaridade (cf. Tabela 2).

Tabela 2 - Propriedades de medidas de similaridade. Adaptado de [3]

Propriedade	Descrição
P1: distância calculada não pode ser negativa: $\boldsymbol{dist(M_0, M_1) \geq 0 \; \forall \; M_0, M_1 \in M}$	Distância de dissimilaridade entre o modelo M_0 e o modelo M_1 deve ser maior ou igual a zero em qualquer modelo pertencente ao conjunto de modelos.
P2: distância calculada é simétrica: $\boldsymbol{dist(M_0, M_1) = dist(M_1, M_0) \; \forall \; M_0, M_1 \in M}$	A distância de dissimilaridade entre o modelo M_0 e o modelo M_1 deve ser igual a distância de dissimilaridade entre o modelo M_1 e o modelo M_0.
P3: distância calculada somente é 0, **se, e somente se,** os modelos de processos são idênticos.	A distância de dissimilaridade entre o modelo M_0 e o modelo M_1 é 0, se, e somente se, o conjunto de atividades e transições contidos em M_0 é equivalente ao conjunto de atividades e transições em M_1.
P4: medida de distância considera as semelhanças e diferenças.	Uma medida de similaridade deve levar em consideração tanto as semelhanças como as diferenças entre dois modelos.
P5: medida de distância mede a similaridade entre atividades.	Uma medida de similaridade deve levar em conta a similaridade das atividades no cálculo da similaridade entre modelos como um todo.
P6: medida de distância é definida para modelos de processos arbitrários.	Medidas de similaridade devem possuir a capacidade de serem aplicadas em qualquer modelo de processo, sem impor restrições, por exemplo: modelos não podem conter *loops*.
P7: medida de distância deve ser computacionalmente eficiente.	Medidas de similaridade devem ser computacionalmente eficiente no cálculo da distância e/ou similaridade de modelos.

Neste estudo foram identificados cerca de 23 relevantes trabalhos relacionadas as medidas de similaridades de modelos de processos. A Tabela 3 apresenta a

aderência das propriedades das medidas de similaridades em cada proposta identificada.

Tabela 3 - Aderência das medidas de similaridade às propriedades de medidas de similaridade de processo. Adaptado de [3].

Medidas	Propriedades						
	P1	P2	P3	P4	P5	P6	P7
Medidas baseadas na correspondência de nós e arestas sem considerar o fluxo de controle							
Similaridade de rótulos [39]	S	S	S	S	S	S	S
Correspondência de rótulos [20]	S	N	S	S	S	S	S
Sintática, estrutural e linguística de rótulos de atividades [44]	S	N	S	S	S	S	S
Estimativa de características [45]	S	S	S	S	S	S	S
Nós e arestas comuns [46]	S	S	N	S	S	S	S
Nós comuns e arestas ponderadas [47]	S	S	N	S	S	S	S
Medidas baseadas na distância de edição de grafos							
Distância de edição de grafos [20]	S	S	S	S	S	S	S
Distância de edição de grafos [41]	S	S	S	N	N	S	S
Distância de edição de grafos e similaridade de rótulos [48]	S	S	S	S	S	S	S
Combinação de distância de edição de texto e distância de edição de grafos [49]	S	S	S	N	S	S	S
Distância de edição de grafos por mudança de operações em alto nível [50]	S	S	S	S	N	N	S
Distância de edição de arvores entre modelos de processos representados como arvores [51]	S	S	N	S	N	S	S
Distância de edição entre modelos reduzidos [10]	S	N	S	S	N	S	S
Medidas que analisam as dependências causais entre as atividades							
Comparação da dependência de grafos [40,52]	S	S	S	S	N	S	S
Comparação da dependência de grafos [9]	S	S	S	S	N	S	S
Medida referencial [43]	S	S	S	S	N	S	N
Relação de transições adjacentes (TAR) [43]	S	S	S	S	N	S	S
Comportamento causal [53]	S	S	S	S	N	N	S
Comportamento causal (*Causal footprints*) [20]	S	S	S	S	N	S	N
Conjunto de traços como *n-grams* [54]	S	N	S	N	N	S	N
Medidas que comparam conjuntos de traços ou *log* de eventos							
Maior subsequência comum nos traços de eventos [55]	S	S	N	S	N	S	N
Baseada nas principais sequencias de transições [56]	S	S	N	S	N	S	N
Baseada em traços de eventos [57]	S	S	N	S	N	S	S

Como já dito anteriormente, estas medidas podem ser segmentadas em 4 grupos de acordo com as técnicas empregadas. As subseções a seguir contextualizam as diferentes técnicas para medir a similaridade entre modelos de processos.

3.1.3 Medidas baseada na correspondência entre nós e arestas

Similaridade de rótulos

Medidas de similaridade baseadas na correspondência entre nós e arestas são consideradas simples e de baixo custo computacional. A ideia principal está em comparar os rótulos das atividades presentes nos nós para determinar o grau de similaridade entre modelos de processo. Neste caso, a ordem de execução das atividades dos modelos de processo não é considerada no cálculo.

Uma abordagem simples para calcular a similaridade de modelos de processo é apresentada por [39], onde considera-se o número de rótulos de atividades equivalentes entre dois modelos de processo. O grau de similaridade dos modelos de processo é definido em duas etapas. Na primeira etapa são identificados os rótulos de atividades comuns entre os processos, e posteriormente é calculada a similaridade de rótulos dos modelos (cf. equação E3).

$$sim(M_0, M_1) = 2 \times \frac{\sum(A_0 \cap A_1)}{(A_0 + A_1)} \quad \text{E3}$$

Onde,

- A_0 representa atividades pertencentes ao conjunto de nós do modelo de processos M_0 e
- A_1 representa atividades pertencentes ao conjunto de nós do modelo de processos M_1.

Correspondência de rótulos

Outra abordagem proposta determina o grau de similaridade entre dois modelos de processo a partir de uma função de correlação de rótulos das atividades para

determinar a equivalência entre modelos [20]. Para calcular o grau de similaridade entre dois modelos, primeiro, define-se um mapeamento das atividades equivalentes nos modelos. A equivalência das atividades é calculada utilizando uma função de correlação $corr(x, map(x))$ dos rótulos individuais x entre os modelos. Somente são adicionadas no mapeamento as atividades que possuírem similaridade maior que um limiar estipulado. Determinado o mapeamento das atividades similares, o grau de similaridade entre dois modelos de processo é calculado (cf. equação E4).

$$sim(M_o, M_1) = \frac{2 \sum_{x \in A_0} corr(x, map(x))}{(\sum N_0 + \sum E_0 \sum N_1 + \sum E_1)} \quad \text{E4}$$

Onde,

- N_0 representa o conjunto de nós do modelo de processos M_0;
- E_0 representa o conjunto de arestas do modelo de processos M_0;
- A_0 representa atividades pertencentes ao conjunto de nós do modelo de processos M_0;
- N_1 representa o conjunto de nós do modelo de processos de M_1; e
- E_1 representa o conjunto de arestas do modelo de processos de M_1.

Estimativa de características

A abordagem apresentada por [45] endereça o problema de recuperação eficiente de modelos de processos similares em relação a um modelo de consulta. O cálculo de similaridade é divido em 3 etapas. Primeiramente, a similaridade das atividades é calculada por meio da distância de edição dos textos dos rótulos das atividades. Em seguida é calculada a similaridade estrutural (funções) de cada nó (cf. Tabela 4), onde são considerados 5 diferentes funções: nós que são início, fim, sequência, divisão ou junção.

Tabela 4 - Funções dos nós no modelo de processo.

Função	Arestas de Entrada	Arestas de Saída
Inicio	0	
Fim		0
Sequência	1	1
Divisão		≥ 2
Junção	≥ 2	

A similaridade estrutural entre dois nós é calculada conforme equações a seguir:

- 1 se ambos os nós possuem função de início e fim;
- $1 - \frac{|succA - succ\ |}{2\ (succA + succ\ )}$ se ambos os nós possuem função de início mas não de fim;
- $1 - \frac{|predA - predB|}{2\ (predA + pred\ )}$ se ambos os nós possuem função de fim mas não início;
- $1 - \frac{|succA - succ\ |}{2\ (succA + succB)} - \frac{|predA - pred\ |}{2\ (predA + pred\ )}$ se ambos os nós possuem função de junção, divisão ou sequência.

Onde,

- *succA* representa o número de nós sucessores de um determinado nó do modelo M_0;
- *succB* representa o número de nós sucessores de um determinado nó do modelo M_1;
- *predA* representa o número de nós predecessores de um determinado nó do modelo M_0 ;e
- *predB* representa o número de nós predecessores de um determinado nó do modelo M_1

O mapeamento entre $n \in N_0$ e $m \in N_1$ é estabelecido se, e somente se, os valores calculados da similaridade de rótulos entre n e m forem maiores que limiares estabelecidos, a saber:

- Rótulos de atividades são semelhantes em um alto grau; ou $lsim(n,m) \geq limiar_{alto}$; ou
- Funções são semelhantes e os rótulos de atividades são semelhantes em médio grau, i.e. $rdsim(n,m) \geq rLimiar$ e $lsim(n,m) \geq limiar_{médio}$.

Por fim, a similaridade entre dois modelos de processo é dada pelo número de nós correspondentes em relação ao número total de nós;

3.1.4 Medidas baseadas na distância de edição de grafos

Distância de edição de grafos

Medidas apresentadas em ([10][48]) visam medir a similaridade entre modelos de processos considerando a estrutura dos modelos comparados, i.e., o quanto os nós e as transições são equivalentes entre os mesmos.

Uma abordagem baseada em redução de grafos é apresentada em [10] e consiste em determinar o grau de similaridade entre modelos de processos por meio do número de operações necessárias de edição de grafos para modificar um modelo ao outro. Nesta abordagem algumas premissas são estabelecidas para determinar o grau de similaridade entre modelos, a saber:

I. Um modelo de processo M_0 é estruturalmente equivalente a outro modelo de processo M_1 quando o número de nós e arestas for igual entre os modelos, onde $N_0 = N_1$ e $E_o = E_1$.
II. Um modelo de processo M_o é dito estruturalmente contido em M_1 quando $N_0 \subseteq N_1$ e N_1 preserva as restrições, i.e., a ordem das atividades.
III. A similaridade entre modelos de processos é dada como 1 se M_0 é equivalente a M_1 ou está contido em M_1.

Para determinar se um modelo é equivalente ou está contido em outro modelo, foi utilizado o método denominado como *SELECTIVE_REDUCE* [11], que aplica

técnicas de redução de grafos para determinar a similaridade entre modelos de processos. O primeiro passo é eliminar todas as atividades de M_1 que não estão contidas em M_0, e em seguida remover todas as arestas que não são necessárias para determinar a ordem das atividades.

Na proposta apresentada originalmente em [12], estas regras visam remover todas as estruturas de um grafo de processo que estão corretas. Contudo, em [11], tais regras são utilizadas para reduzir um modelo M_1 para algo que seja equivalente ou esteja contido em M_o. Por fim, o grau de similaridade entre dois modelos considera apenas o número de arestas comuns entre os modelos [10], sendo definido pela equação E5:

$$sim(M_0, M_1) = \frac{\sum(E_1^{red} \cap E_0)}{\sum E_1^{red}} \quad \text{E5}$$

Onde,

- E_0 é conjunto de arrestas do modelo de processo M_0 e
- $\boldsymbol{E_1^{red}}$ é o conjunto reduzido de arrestas do modelo de processos M_1.

Distância de edição de grafos e similaridade de rótulos

Uma abordagem para fazer a junção de dois modelos de processos similares foi proposta em [48], visando consolidar e eliminar as redundâncias de modelos de processos entre companhias, e consiste, primeiramente, em calcular a similaridade entre modelos de processos para posteriormente sugerir a junção dos mesmos. A definição do grau de similaridade entre modelos de processos é dividida em três etapas.

A *primeira etapa* é responsável por realizar o mapeamento entre os pares de nós de dois modelos de processos. Aqui, deve-se considerar que nós de diferentes tipos não podem ser mapeados. O grau de similaridade do mapeamento entre nós é calculado por meio da similaridade sintática dos seus rótulos vis-à-vis a distância de edição de textos e similaridade linguística.

A *segunda etapa* é responsável por colocar em prática algumas métricas que avaliam o contexto dos nós dos modelos de processos:

- *Substituições de nós*: um nó em um modelo é substituído por um outro nó em um outro modelo, se e somente se, eles aparecem no mapeamento.
- *Inserções e remoções de nós*: um nó é inserido ou removido de um modelo se eles não aparecem no mapeamento.
- *Substituições de arestas*: uma aresta de um respectivo nó é substituída por uma outra aresta de outro modelo se ambos nós, origem e destino, forem iguais.
- *Inserções e remoções de arestas*: uma aresta é inserida ou removida de um modelo se a mesma não foi substituída.

Por fim, a *última etapa* consiste em calcular a similaridade entre dois modelos de processos utilizando o grau de similaridade obtido na *primeira etapa* e a informação sobre os nós e as arestas substituídas, inseridas e removidas da *segunda etapa*. A medida utiliza para calcular a similaridade entre dois modelos é definida pela equação E6.

$$1.0 - \frac{wskipn \times fskipn + wskipe \times fskipe + wsubn \times fsubn}{wskipn + wskipe + wsubn} \quad \text{E6}$$

Onde $wsubn$ representa o peso atribuído aos nós substituídos, $wskipn$ aos pesos dos nós inseridos e removidos e $wskipe$ os pesos das arestas inseridas e removidas. As frações dos nós inseridos/removidos, das arestas inseridas/removidas e da média da distância dos nós substituídos são denotadas por $fskipn$, $fskipe$ e $fsubn$ respectivamente.

3.1.5 Medidas de dependências causais entre as atividades

Comparação da dependência de grafos

Medidas de similaridade de dependências causais entre atividades visam aferir a similaridade entre dois modelos de processos de acordo com as relações entre atividades e transições. A grande diferença de tais medidas em relação as que consideram somente a estrutura dos modelos, está relacionada ao fato que em alguma medida as relações adjacentes, implícitas e diretas entre os nós são consideradas [20].

Neste contexto, há uma abordagem particular onde a representação das dependências causais entre as atividades é definida por dois vetores [9]:

i. *Vetor de atividades*: provê as probabilidades de execução de cada atividade A_0 do modelo M_o em A_1 do modelo M_1;
ii. *Vetor de transições*: provê o resultado da multiplicação da probabilidade de execução de cada atividade pela distância entre duas atividades—a distância entre duas atividades é proporcional ao número de transições entre as mesmas.

A Figura 17 ilustra o processo de transformação de uma sequência de atividades com transições implícitas em um grafo completo ponderado de dependência (em inglês, wCDG—*weighted Complete Dependency Graph*). O valor atribuído as transições entre as atividades são definidas pelo inverso da distância entre uma transição direta de duas atividades, de modo que o peso para transições diretas é 1 e o peso para uma transição implícita é um valor entre [0, 1]. Os vetores de atividades e transições de um modelo de processo são ilustrados na Figura 18.

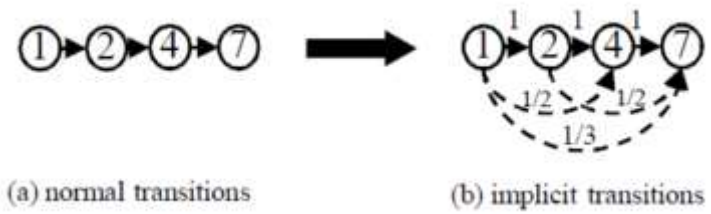
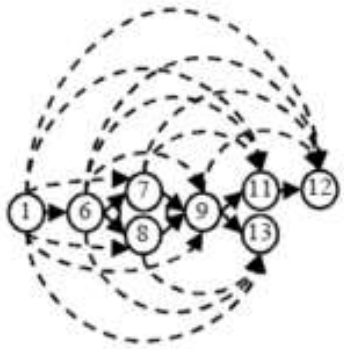


Figura 17 - Transformação de um modelo processo para wCDG. Fonte: [9].



Activity vector $a_1=(1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0.5, 0.5, 0.5)$

Transition vector $t_1=(t_{ij})$, *where*

$t_{1,6}=t_{6,7}=t_{6,8}=t_{7,9}=t_{8,9}=1$
$t_{1,7}=t_{1,8}=t_{6,9}=t_{9,11}=t_{8,13}=1/2$
$t_{1,9}=1/3, t_{7,11}=t_{7,13}=t_{8,11}=t_{8,13}=t_{11,12}=t_{9,12}=1/4$
$t_{6,11}=t_{6,13}=t_{7,12}=t_{8,12}=1/6$
$t_{1,11}=t_{1,13}=t_{6,12}=1/8$
$t_{1,12}=1/10$
other $t_{ij}=0$

Figura 18 - Exemplo dos vetores do modelo de processo. Fonte: [9].

Após a construção dos vetores das atividades e das transições, a similaridade dos modelos de processo é dada pelo coeficiente de cosseno. Este último permite obter um valor alto para os dois vetores que possuem elementos em comum com valores altos. As medidas de similaridade para os vetores de atividades e transições são definidas pelas equações E7 e E8, respectivamente.

$$sim_{act}(M_o, M_1) = \frac{A_0 \times A_1}{|A_o||A_1|} = \frac{\sum A_{i,0}A_{i,1}}{\sqrt{\sum A_{i,0}^2 \sum A_{i,1}^2}} \quad \text{E7}$$

$$sim_{trans}(M_o, M_1) = \frac{E_0 \times E_1}{|E_0||E_0|} = \frac{\sum E_{i,0}E_{i,1}}{\sqrt{\sum E_{i,0}^2 \sum E_{i,1}^2}} \quad \text{E8}$$

Por fim, a similaridade total de dois modelos de processos é obtida por meio da soma dos pesos dos vetores de atividades e de transições, sendo definida pela equação F9.

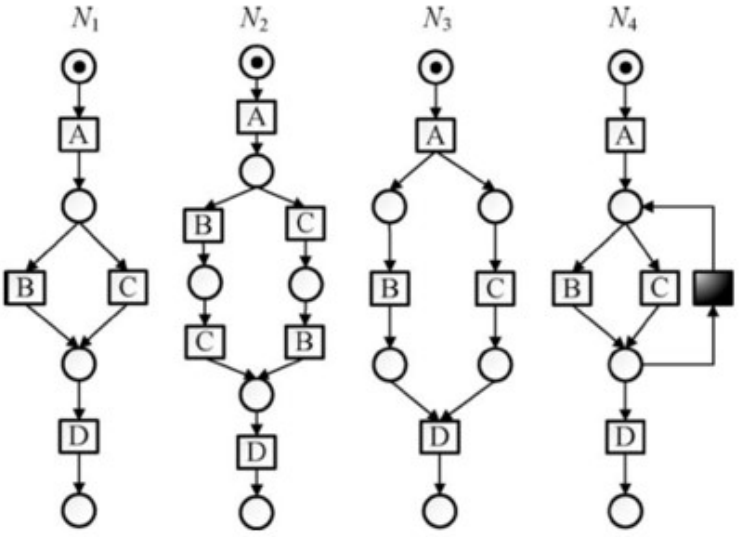
$$sim(M_0, M_1) = \alpha sim_{act}(M_0, M_1) + (1 - \alpha) sim_{trans}(M_0, M_1) \quad \text{E9}$$

A abordagem apresentada em [9] não esclarece como transformar um modelo de processo que contem laços nos vetores de atividades e transições, sendo esta

considerada uma limitação [3], de tal forma que é necessário remover os laços presentes nos modelos de processos a serem comparados.

Relações de adjacência de transições

Outra proposta que analisa as dependências causais entre as atividades com o objetivo de comparar o comportamento dos modelos de processos foi apresentada por [43]. Esta proposta é baseada nas relações de adjacência de transições (TAR) dos modelos de processos. Como um conjunto de TAR descreve a ordem das transições que aparecem em todas as sequências de disparos possíveis de um modelo representado por uma rede de Petri, considera-se que o comportamento de um modelo de processo pode ser especificado. A Figura 19 ilustra quatro modelos de processos N1, N2, N3 e N4 e os conjuntos de TAR: {AB; AC; BD; CD}, {AB; AC; BC; CB; BD; CD}, {AB; AC; BD; CD} e {AB; AC; BC; CB; BB; CC; BD; CD} respectivos para cada modelo de processo.



N_1	A	B	C	D
A		1	1	
B				1
C				1
D				

N_2	A	B	C	D
A		1	1	
B			1	1
C		1		1
D				

N_3	A	B	C	D
A		1	1	
B				1
C				1
D				

N_4	A	B	C	D
A		1	1	
B		1	1	1
C		1	1	1
D				

Figura 19 - Conjunto de TAR dos modelos de processos. Adaptado de [43].

Baseado nos conjuntos de relações adjacentes, a similaridade entre dois modelos de processos é dada por meio da seguinte fórmula:

$$sim(M_o, M_1) = \frac{\sum(TAR_o \cap TAR_1)}{\sum(TAR_o \cup TAR_1)} \quad \text{F10}$$

Uma das vantagens apresentadas por [43] na expressão do comportamento dos modelos de processos por meio das relações de transições adjacentes é o menor custo computacional quando comparado a construção completa das sequências de disparo. Ressalta-se a possibilidade de aplicar esta medida em aplicações como: agrupamento de modelos de processos, mineração de processos, integração e recuperação de modelos de processos [43].

3.2 Considerações sobre o Capítulo

Este capítulo apresentou uma revisão da literatura mais abrangente sobre propriedades e medidas de similaridade. Foram contextualizadas as propriedades que as medidas de similaridade devem apresentar aderência, sendo estas utilizadas como critérios de seleção de medidas. Também foram contextualizadas medidas de similaridade que possuíam maior aderência face as propriedades de medidas de similaridade. Estas representaram diferentes técnicas utilizadas para avaliar o grau de correspondência entre modelos de processos: (1) Medidas baseadas na correspondência entre nós e arestas; (2) Medidas baseadas na distância de edição de grafos; (3) Medidas baseadas nas dependências causais entre as atividades.

CAPÍTULO 4 - ESTRUTURAÇÃO DA PESQUISA

Este capítulo apresenta a abordagem metodológica colocada em prática para o desenvolvimento da pesquisa. Nesta linha serão usadas algumas referências para estabelecer os conceitos necessários. Na sequência será caracterizada a forma de trabalho e definida a estratégia, assim como as etapas utilizadas para a consecução dos objetivos previamente enumerados.

4.1 Método de Pesquisa

A apresentação e interpretação dos fenômenos estudados será caracterizada por um experimento. O objetivo de um experimento é coletar dados suficientes de uma população—todos de um mesmo ambiente—a fim de obter um resultado estatisticamente significativo sobre um atributo estudado em comparação com outro atributo [26]. A Figura 20 encerra as fases do experimento.

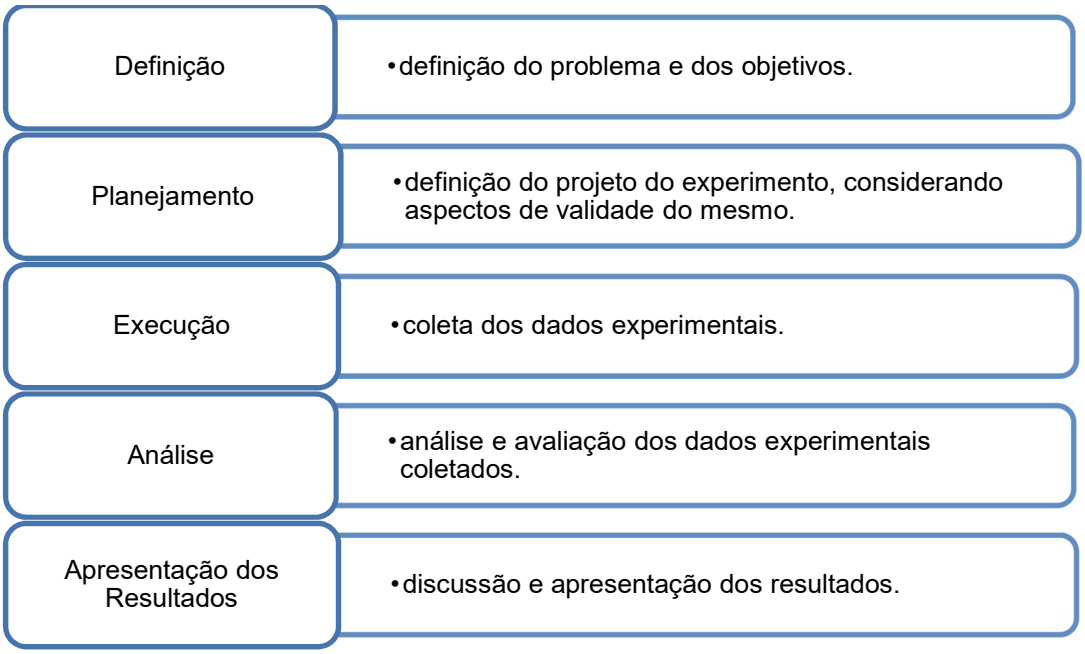


Figura 20 - Fases do experimento. Fonte: [26].

Definida as fases do método escolhido, a seguir são apresentadas as questões de pesquisa.

4.2 Questões de Pesquisa

O objetivo geral desta pesquisa é analisar se diferentes medidas de similaridade de modelos de processos impactam na qualidade do agrupamento de instâncias de processo. Tal análise visa estabelecer medidas de similaridade mais apropriadas para o agrupamento de instâncias quando aplicadas em conjunto com a Mineração de Processo. Neste contexto, as seguintes questões de pesquisa foram estabelecidas:

RQ01: As diferentes medidas de similaridade impactam na qualidade do agrupamento de instâncias de processo, e consequentemente no resultado da Mineração de Processo?

RQ02: É possível estabelecer medidas de similaridade mais apropriadas para o agrupamento de processos em processos não estruturados, em particular, quando aplicadas em conjunto com a Mineração de Processo?

4.3 Estratégia de Pesquisa

Esta seção descreve a estratégia da pesquisa, resumida de forma gráfica cf. Figura 21, cujo resultado é o projeto do experimento pronto para a execução. Esta pesquisa foi dividida em duas fases: (1) Estudo exploratório para identificar e selecionar medidas de similaridade aplicáveis no processo de agrupamento de instâncias de processo; e (2) Projeto do experimento para definir o problema de pesquisa, desenhar e construir o experimento, coletar dados e analisa-los a fim de responder as questões de pesquisa já relatadas.

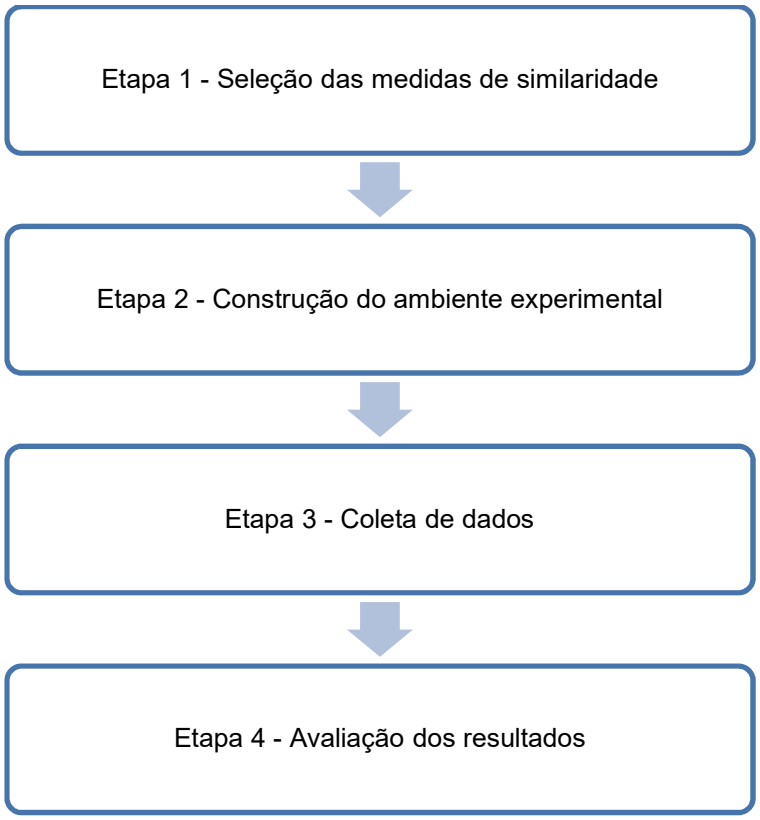


Figura 21- Etapas do planejamento do experimento. Fonte: o Autor, 2017

Nas subseções a seguir serão detalhadas cada etapa do experimento.

4.3.1 Etapa 1 - Identificação e seleção de medidas de similaridade

Esta etapa visa identificar e selecionar diferentes técnicas de medidas de similaridade aplicadas a modelos de processos, na forma como elas foram contextualizadas anteriormente na revisão da literatura.

Deve-se destacar, a partir da contextualização de tais técnicas de medição, enumeração de diferentes propriedades de medidas de similaridade (cf. Tabela 2), assim como 23 propostas de medição da similaridade entre modelos de processos. Estas foram segregadas em 4 diferentes grupos de acordo com as técnicas empregadas (cf. Tabela 3). E baseado nas propriedades das medidas de similaridade, os seguintes critérios de exclusão foram utilizados:

- (EC.1) Distância pode ser negativa;
- (EC.2) Distância calculada não é simétrica;

- (EC.3) Distância calculada somente é 0, se, e somente se, os modelos de processos são idênticos.
- (EC.4) Medida de distância não considera as semelhanças e diferenças;
- (EC.5) Medida de distância não é definida para modelos de processos arbitrários; e
- (EC.6) Medida de distância não é computacionalmente eficiente.

Como neste estudo os modelos de processo são representados por meio de uma rede de Petri, as técnicas para medir a similaridade de modelos devem possibilitar o cálculo neste tipo de representação. Desta forma, o critério apresentado EC.7 é estabelecido:

- (EC.7) Medida não é aplicada a modelos representados por rede de Petri.

Por fim, nos casos de abordagens que empregam técnicas similares, somente será escolhida uma, visto que o objetivo do trabalho não é explorar as particularidades de cada proposta. Com isso, o último critério de exclusão é estabelecido:

- (EC.8) Medida é similar em mais de uma proposta.

Dado o estabelecimento dos critérios de exclusão, eles foram aplicados nas medidas de similaridade apresentadas na Tabela 3, com o objetivo de melhor definir o escopo de trabalho. Neste sentido, a Figura 22 mostra o processo de seleção das medidas de similaridade.

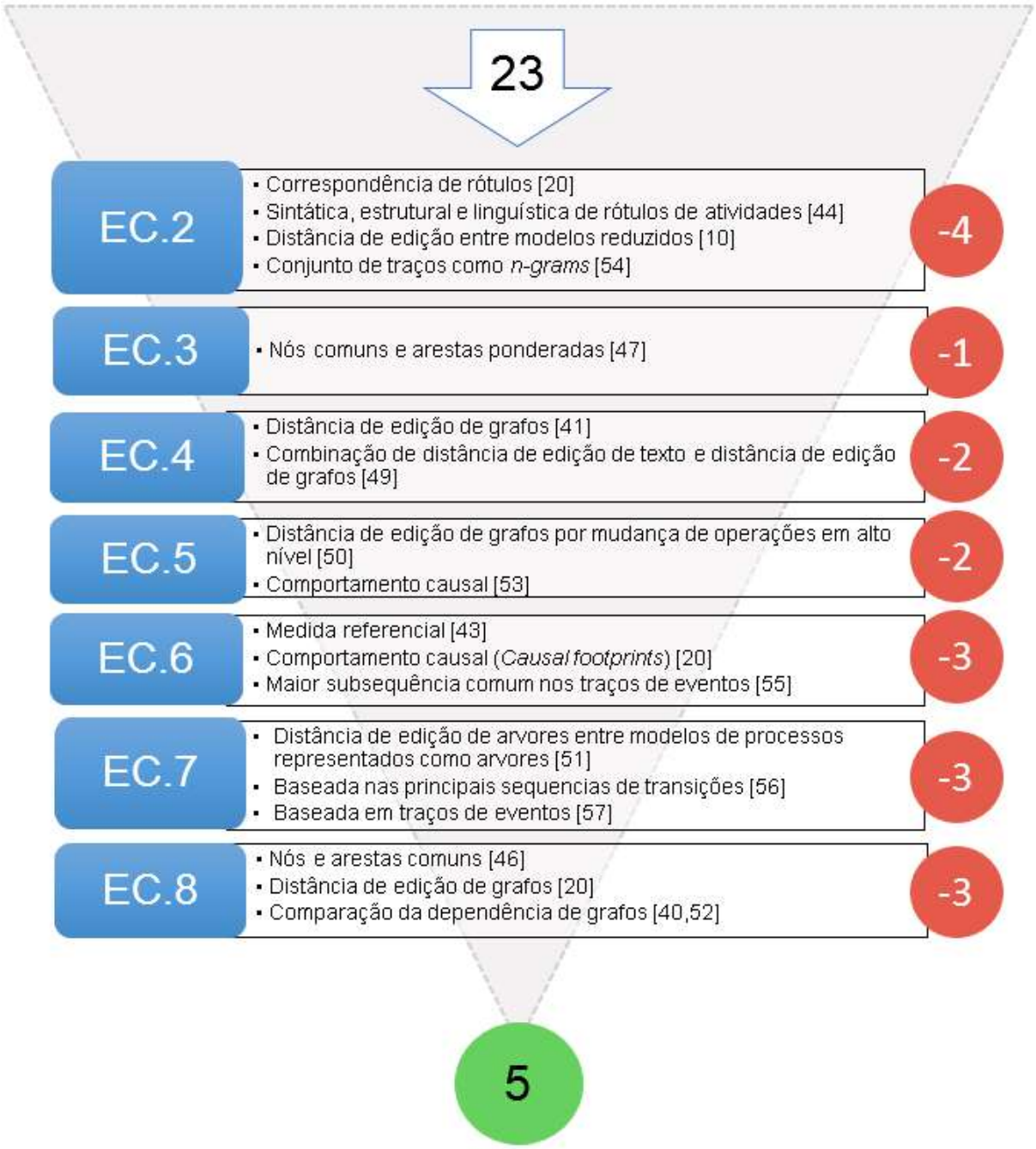


Figura 22 - Processo de seleção de medidas de similaridade de instâncias de processos.

Das 23 abordagens elencadas, foram selecionadas para o experimento 5 delas, ressaltando 3 diferentes técnicas para calcular a similaridade entre modelos de processo (cf. Tabela 5).

Tabela 5 - Medidas de similaridade selecionadas para o experimento.

M1 - Similaridade de rótulos [39];
M2 - Comparação da dependência de grafos [9];
M3 - Estimativa de características [45];
M4 - Distância da edição de grafos e similaridade de rótulos [48];
M5 - Relação de transições adjacentes (TAR) [43];

4.3.2 Etapa 2 – Construção do ambiente experimental

Esta etapa consiste em construir o ambiente experimental para coleta de dados. Neste sentido, fez-se necessário implementar um algoritmo de agrupamento incremental e em conjunto aplicar as métricas de similaridade selecionadas.

Antes de descrever o algoritmo de agrupamento utilizado, serão reforçados alguns conceitos e notações importantes utilizados neste estudo. Um *log* de eventos (L) contém um conjunto de traços. Um traço (T) é uma sequência finita de eventos, e ela descreve uma instância de processo em específico. Uma instância de processo reflete o comportamento real de um modelo de processo (PM), sendo representado aqui por uma *Rede de Petri*. Para transformar uma instância de processo em um modelo de processo é necessário utilizar alguma técnica de descoberta de processo, neste caso foi selecionado o *Inductive Miner*. Um grupo de processos (C) é composto por um modelo de processo e um conjunto de traços. Por fim, a saída do algoritmo de agrupamento definido a seguir é um conjunto de agrupamentos (CS).

O algoritmo utilizado é *particional, exclusivo e completo* e realiza agrupamentos incrementais—maiores detalhes em [37]. A ideia básica do algoritmo (cf. Figura 23) é transformar cada traço de eventos em um modelo de processo para medir a similaridade entre elementos. Um limiar de similaridade é utilizado para determinar se um novo agrupamento é criado com o modelo de processo do traço atual, ou se o traço atual é adicionado à um agrupamento já existente. Quando um traço é adicionado para um agrupamento já existente o centroide que representa o modelo de processo não é recalculado, permanecendo o primeiro modelo de processo adicionado. Isso significa que o cálculo da similaridade entre instâncias de processos, somente será realizado utilizando o primeiro modelo adicionado de cada

agrupamento. De forma geral, a ordem de entrada das instâncias de processos pode ocasionar uma variação nos resultados obtidos.

Algorithm 1: Incremental trace clustering

```
input  : Event Log L, Threshold
output: Set of clusters CS
Let T denote a trace;
Let PM denote a process model;
Let C denote a cluster composed by a set of T and a PM;
Let CA denote a candidate cluster;
foreach T ⊂ L do
    PM ← mine(T)
    if CS = ∅ then
        Add cluster C to CS with C ← (PM, T) ;
    else
        foreach C ∈ CS do
            Calculate similarity S between PM ⊂ C and PM;
            if S > Threshold then
                Add C to candidate list of clusters CA;
            end
        end
        if CA = ∅ then
            Add new cluster C to CS with C ← (PM, T) ;
        else
            Get cluster C that contains the max similarity
             C ← max(S, CA) ;
            Add T into cluster C with C ← T ;
        end
    end
end
```

Figura 23 - Algoritmo incremental de agrupamento de traços.

As métricas de similaridade de modelos de processos, conforme critérios contextualizados na seção anterior, foram aplicadas na etapa responsável por calcular a distância entre os centroides existentes e o modelo de processo descoberto em cada iteração no *log* de eventos. Ao final do processo de agrupamento, o resultado é um conjunto de agrupamentos com traços (i.e., instâncias de processo) similares.

4.3.3 Etapa 3 – Coleta de dados

Nesta etapa foi conduzida a coleta de dados, onde o algoritmo de agrupamento incremental foi aplicado em 8 *logs* de eventos (cf. Tabela 6), sendo 2 *logs* de eventos sintéticos, e 6 *logs* de eventos extraídos de processos reais. Estes *logs* de eventos foram utilizados para *benchmark* em diversos estudos na área de Mineração de Processos.

Tabela 6 – *Log* de eventos de benchmark selecionados para o experimento.

Segmento	Processo	Casos	Eventos	Publicado	Fonte
Log Sintético	Processo com 0% de ruído	100.000	894.708	2017	[60]
Log Sintético	Processo com 30% de ruído	100.000	894.708	2017	[60]
Saúde	Tratamento de pacientes	1.143	150.291	2011	[61]
Saúde	Faturamento de serviços médios	100.000	451.359	2011	[61]
Público	Recebimento de pedidos de contrução	1434	8.577	2014	[62]
Saúde	Tratamento de pacientes	1.050	15.214	2016	[63]
Financeiro	Pedido de Empréstimo	13087	262.200	2012	[64]
Público	Gerenciamento de multas	150.370	561.470	2015	[65]

O processo de coleta de dados é dividido em 3 etapas (cf. Figura 24):

I. Em cada *log* de eventos selecionado foi aplicado o algoritmo de agrupamento de instâncias de processos (cf. Figura 23) com as medidas de similaridades selecionadas (cf. Tabela 5) e com diferentes limiares de similaridade: 40%, 60% e 80%.

II. Em cada execução do algoritmo de agrupamento, gerou-se um conjunto de agrupamentos com tamanho igual ou superior a 1. Cada agrupamento contém um conjunto de instâncias de processos. Foi aplicada a técnica de descoberta de modelos de processos *Inductive Miner* para descobrir o modelo de processo que representava cada agrupamento.

III. Baseado no modelo de processo descoberto de cada agrupamento foi calculado a qualidade do modelo por meio de três métricas: *Recall / Fitness* [38], *Precision* [34] e *F1 Score*[6].

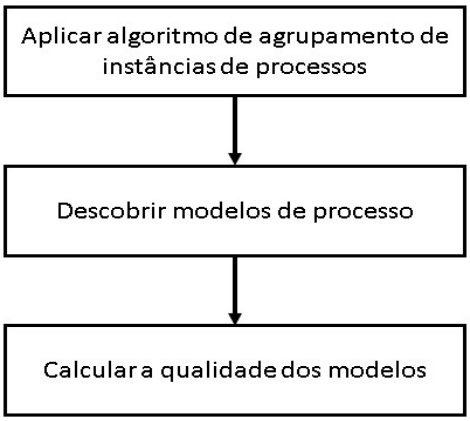


Figura 24 - Processo da coleta de dados

A qualidade calculada dos modelos gerados pelo algoritmo de agrupamento de instâncias de processos com os diferentes limiares de similaridade e as medidas de similaridade selecionadas foram analisados conforme critérios estabelecidos na seção a seguir, com o objetivo de responder as questões de pesquisa. A Figura 25 apresenta a amostragem da coleta de dados realizada no experimento.

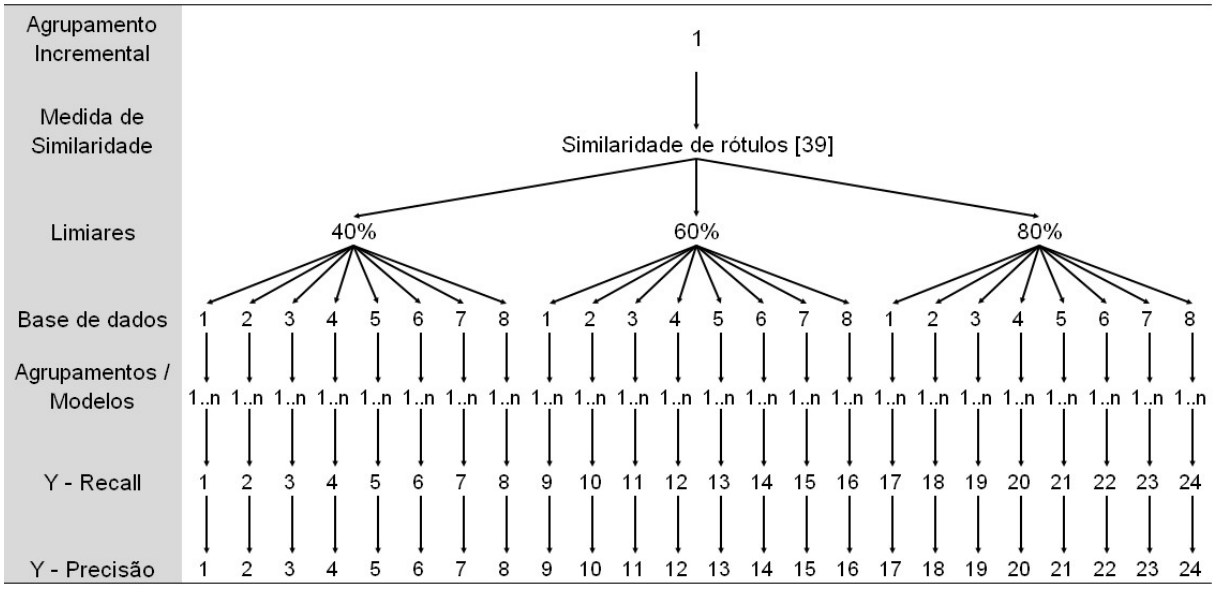


Figura 25 - Esquema de amostragem da coleta de dados.

4.3.4 Etapa 4 – Avaliação dos resultados

As análises dos dados coletados foram divididas em duas etapas. A primeira etapa fundamenta-se no método estatístico não paramétrico de *Kruskal-Wallis* para verificar a hipótese de que a qualidade dos modelos de processos representados pelos agrupamentos pode diferenciar-se *vis-à-vis* as medidas de similaridade. Como os conjuntos de agrupamentos resultantes da aplicação do algortimo de agrupamento de instâncias de processos com diferentes medidas de similaridade podem gerar conjuntos com tamanhos diferentes, este teste foi o mais adequado, pois sua aplicação é recomendado quando há três ou mais situações experimentais com amostras não relacionadas.

Na segunda etapa, caso se verifique que há diferença estatística na qualidade dos modelos de processos face as medidas de similaridade, aplicar-se-á o teste *post-hoc de Dunn-Bonferroni* com o objetivo de identificar quais medidas de similaridade possuem diferenças estatísticas entre si.

A partir destas duas etapas é possível responder à questão de pesquisa RQ01, pois elas permitem verificar se diferentes medidas de similaridade impactam na qualidade do agrupamento de instâncias de processo, e consequentemente no resultado da Mineração de Processo, assim como a questão de pesquisa RQ02, pois elas também permitem verificar se é possível estabelecer medidas de similaridade mais apropriadas para o agrupamento de processos, em particular, quando aplicadas em conjunto com a Mineração de Processo.

4.4 Considerações sobre o Capítulo

Este capítulo apresentou o planejamento e execução do experimento proposto para avaliar as medidas de similaridade aplicadas no agrupamento de instâncias de processo face a qualidade dos modelos de processos gerados.

O experimento foi realizado seguindo o método experimental, utilizando um estudo em duas etapas: (1) Identificação e seleção de medidas de similaridade; (2) Realização do experimento, a partir do qual as qualidades dos modelos gerados foram

coletadas face as diferentes medidas de similaridade aplicadas no agrupamento incremental de instâncias de processos. Por último, foi apresentado como os dados coletados foram analisados, considerando os testes estatísticos utilizados para responder as questões de pesquisas elaboradas.

CAPÍTULO 5 - ANÁLISE E DISCUSSÃO DOS RESULTADOS

Esta seção apresenta a análise consolidada dos resultados obtidos com base nos experimentos descritos e colocados em prática anteriormente. Neste trabalho, analisou-se os resultados obtidos a partir das execuções do processo de agrupamento incremental com as medidas de similaridade selecionadas anteriormente em oito bases de dados, sendo duas sintéticas e seis provenientes de processos reais. As tabelas apresentadas a seguir denotam o número de agrupamentos criados (n), a média (*mean*), desvio padrão (*sd*), mínimo (min), mediana (*med*) e valor máximo (*max*) para cada uma das métricas de qualidade. As métricas de qualidade são representadas por:

- *Recall* também conhecido como *tp-rate*. Ela mede a proporção de instâncias positivas classificadas como positivas (tp / p). Os traços no *log* de eventos são instâncias positivas e quando uma instância pode ser reproduzida pelo modelo, então a instância é de fato classificada como positiva. Neste contexto, a medida *Recall* representa o *Fitness* do modelo descoberto [6].
- Precisão remete ao comportamento presente no modelo de processo. Um modelo que não é preciso é *underfitting.* E *underfitting* refere-se ao problema de modelos muitos generalistas que permitem comportamentos muito diferentes do presente no *log* de eventos [6].
- *F1 Score* representa a média harmônica entre Precisão e *Recall*: (2 × *Precision* × *Recall*) / (*Precision* + *Recall*). Se a precisão e *Recall* são ruins (i.e., próximos de 0), então *F1 Score* também será próximo de 0. Apenas se *Precisão* e *Recall* são boas, o valor de *F1 Score* será próximo de 1.

Nas subseções a seguir são apresentados os resultados da execução das medidas de similaridade no agrupamento de instâncias de processo em diferentes bases de dados, considerando diferentes limiares de similaridade.

5.1 *Log* sem ruídos

Esta seção irá apresentar os resultados do experimento aplicado no log de eventos gerados artificialmente a partir da ferramenta *CPN Tools*[2] que representa um processo semiestruturado com 100.000 casos, 894.708 eventos, 8 atividades e 346 diferentes fluxos no processo. Deve-se frisar que um *log* de eventos é composto por instâncias de processos com caminhos infrequentes gerados de forma aleatória.

5.1.1 Análise estatística

A partir dos resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 40% e do teste de *Kruskal Wallis* observa-se que: **(1)** Não existe diferença estatística nos resultados de *Recall*, *Precision* e *F1 Score* (cf. Tabela 7) dentre as medidas de similaridade utilizadas no agrupamento de instâncias de processos.

Tabela 7 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos com limiar de similaridade de 40% para o log de eventos sintético com percentual zero de ruído.

Medida de Similaridade		n	Recall					Precision					FScore				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	1.00	0.41	NA	0.41	0.41	0.41	**0.92**	NA	0.92	**0.92**	0.92	0.56	NA	0.56	0.56	0.56
M2	Comparação da dependência de grafos [9]	1.00	0.41	NA	0.41	0.41	0.41	**0.92**	NA	0.92	**0.92**	0.92	0.56	NA	0.56	0.56	0.56
M3	Estimativa de características [45]	1.00	0.41	NA	0.41	0.41	0.41	**0.92**	NA	0.92	**0.92**	0.92	0.56	NA	0.56	0.56	0.56
M4	Distância da edição de grafos e similaridade de rótulos [48]	1.00	0.41	NA	0.41	0.41	0.41	**0.92**	NA	0.92	**0.92**	0.92	0.56	NA	0.56	0.56	0.56
M5	Relação de transições adjacentes [43]	15.00	**0.75**	0.14	0.58	**0.69**	1.00	0.80	0.23	0.30	**0.92**	1.00	**0.74**	0.14	0.46	**0.76**	1.00
Kruskal-Wallis			p-value = 0.05914					p-value = 0.9998					p-value = 0.2995				

A Tabela 8 e Tabela 9 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 60%, em conjunto com os testes de *Kruskal Wallis* e teste *post-hoc* de *Dunn-Bonferroni* respectivamente. A partir destes resultados, observa-se que: **(1)** a qualidade dos modelos gerados foi superior quando comparado ao processo de agrupamento com limiar de similaridade 40%. (cf.

[2] Ferramenta para editar, simular e analisar Rede de Petri coloridas. Disponível em:http://cpntools.org/

Tabela 8) **(2)** Não há diferença estatística dos resultados de *Recall, Precision* e *F1 Score*. (cf. Tabela 8 e Tabela 9).

Tabela 8 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 60%, log de eventos sintético e percentual de ruído zero.

Medida de Similaridade		n	Recall					Precision					FScore				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	2	**0.852**	0.209	0.704	0.852	1.000	**0.969**	0.044	0.937	0.969	1.000	**0.902**	0.139	0.804	0.902	1.000
M2	Comparação da dependência de grafos [9]	2	**0.852**	0.209	0.704	0.852	1.000	**0.969**	0.044	0.937	0.969	1.000	**0.902**	0.139	0.804	0.902	1.000
M3	Estimativa de características [45]	2	**0.852**	0.209	0.704	0.852	1.000	**0.969**	0.044	0.937	0.969	1.000	**0.902**	0.139	0.804	0.902	1.000
M4	Distância da edição de grafos e similaridade de rótulos [48]	1	0.407	NA	0.407	0.407	0.407	**0.918**	NA	0.918	0.918	0.918	0.564	NA	0.564	0.564	0.564
M5	Relação de transições adjacentes [43]	33	0.764	0.116	0.620	0.706	1.000	0.756	0.224	0.302	0.867	1.000	0.724	0.123	0.446	0.755	1.000
Kruskal-Wallis			p-value = 0.2318					p-value = 0.01586					p-value = 0.01005				

Tabela 9 – Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 60%, *log* de eventos sintético e percentual de ruído zero.

Comparação	p-value	
	Precision	Fscore
M1 - M2	1	1
M1 - M3	1	1
M2 - M3	1	1
M1 - M4	0.7983357	0.10542558
M2 - M4	0.9313916	0.15813838
M3 - M4	1	0.31627675
M1 - M5	0.1188533	0.06515395
M2 - M5	0.1782799	0.07818474
M3 - M5	0.3565598	0.09773093
M4 - M5	1	0.37966109

A Tabela 10 e Tabela 11 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 80%, em conjunto com os testes de *Kruskal Wallis* e teste *post-hoc* de *Dunn-Bonferroni* respectivamente. A partir destes resultados observa-se que: **(1)** Não há diferença estatística na qualidade média de *Recall*, *Precision* e *F1 Score* (cf. e Tabela 11) dentre as medidas de similaridade utilizadas no agrupamento de instâncias de processos.

Tabela 10 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 80%, *log* de eventos sintético e percentual de ruído zero.

Medida de Similaridade		n	Recall					Precision					FScore				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	2	**0.852**	0.209	0.704	0.852	1.000	**0.969**	0.044	0.937	**0.969**	1.000	**0.902**	0.139	0.804	0.902	1.000
M2	Comparação da dependência de grafos [9]	12	0.716	0.120	0.588	0.698	1.000	0.837	0.199	0.302	0.937	1.000	0.751	0.131	0.446	0.767	1.000
M3	Estimativa de características [45]	2	**0.852**	0.209	0.704	0.852	1.000	**0.969**	0.044	0.937	**0.969**	1.000	**0.902**	0.139	0.804	0.902	1.000
M4	Distância da edição de grafos e similaridade de rótulos [48]	8	0.735	0.139	0.582	0.700	1.000	0.898	0.121	0.608	0.937	1.000	0.799	0.103	0.669	0.783	1.000
M5	Relação de transições adjacentes [43]	77	0.794	0.087	0.669	0.794	1.000	0.661	0.227	0.159	0.635	1.000	0.689	0.136	0.267	0.728	1.000
Kruskal-Wallis			p-value = **0.04216**					p-value = **2.018e-05**					p-value = **0.01579**				

Tabela 11 - Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 80%, *log* de eventos sintético e percentual de ruído zero.

Comparação	p-value		
	Recall	Precision	Fscore
M1 - M2	0.30664733	0.5146863	0.3151341
M1 - M3	1	1	1
M2 - M3	0.40886311	0.617623559	1
M1 - M4	0.36925124	0.637252133	0.4124997
M2 - M4	0.71259384	0.66653997	0.5532978
M3 - M4	0.44310149	0.71690865	0.4714282
M1 - M5	0.77658401	**0.042871436**	0.1896419
M2 - M5	0.07842185	**0.013674118**	0.3158373
M3 - M5	0.88752458	0.057161915	0.3792837
M4 - M5	0.49574178	**0.006793633**	0.1791404

5.1.2 Síntese dos resultados

O processo de agrupamento de instâncias de processos foi aplicado no *log* de eventos sintético que caracteriza um processo semiestruturado. No processo de agrupamento com limiar de similaridade 40%, observou-se que a utilização de medidas de similaridade que consideram os rótulos das atividades (M1 e M3), estrutura dos modelos (M4) e relação diretas e indiretas entre as atividades (M2) não geraram agrupamentos de instâncias de processos, permanecendo o *log de eventos*. Esse comportamento explica-se devido à natureza do *log* de eventos gerado, pois grande parte dos casos continham as mesmas atividades, mas com repetições de atividades no mesmo caso. Em contrapartida, a utilização da medida de similaridade que considera a relação das transições adjacentes (M5) resultou no maior número de

agrupamentos devido ela ser sensível as transições existentes nas instâncias de processos. No agrupamento com limiar de similaridade 60%, observou-se que o número de agrupamentos gerados foi maior quando comparado ao limiar 40%, e a qualidade dos modelos gerados foi consideravelmente superior.

Por fim, a partir dos resultados obtidos, não foi possível verificar estatisticamente qual medida obteve o melhor desempenho no agrupamento de instâncias de processos, visto que, não houve diferença estatística dos resultados dentre as medidas de similaridade.

5.2 *Log* com percentual de ruído de 30%

Esta seção irá apresentar os resultados do experimento aplicado no log de eventos gerado artificialmente com a ferramenta *CPN Tools* que representa um processo não estruturado com 100.000 casos, 924.329 eventos, 8 atividades e 2946 diferentes fluxos no processo. Aqui, o *log* de eventos é composto por instâncias de processos com caminhos infrequentes gerados de forma aleatória com a adição de eventos nas instâncias de processos. O nível de ruído adicionado é cerca de 30% em relação ao *log* de eventos com 0% de ruído.

5.2.1 Análise estatística

A partir dos resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 40% e do teste de *Kruskal Wallis* observa-se que: **(1)** Não há diferença estatística nos resultados de *Recall*, *Precision* e *F1 Score* (cf. Tabela 12) dentre as medidas de similaridade utilizadas no agrupamento de instâncias de processos. **(2)** M5 resultou no maior número de agrupamentos gerados no agrupamento de instâncias de processos.

Tabela 12- Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 40%, *log* de eventos sintético e percentual de ruído 30%.

Medida de Similaridade		n	Recall					Precision					FScore				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	1.00	0.282	NA	0.282	0.282	0.282	0.625	NA	0.625	**0.625**	0.625	0.388	NA	0.388	0.388	0.388
M2	Comparação da dependência de grafos [9]	2.00	0.452	0.175	0.328	0.452	0.576	0.549	0.108	0.472	**0.549**	0.625	0.475	0.063	0.430	0.475	0.519
M3	Estimativa de características [45]	1.00	0.282	NA	0.282	0.282	0.282	0.625	NA	0.625	**0.625**	0.625	0.388	NA	0.388	0.388	0.388
M4	Distância da edição de grafos e similaridade de rótulos [48]	1.00	0.282	NA	0.282	0.282	0.282	0.625	NA	0.625	**0.625**	0.625	0.388	NA	0.388	0.388	0.388
M5	Relação de transições adjacentes [43]	24.00	**0.632**	0.125	0.471	0.592	0.923	**0.642**	0.188	0.029	0.675	1.000	**0.607**	0.151	0.055	0.622	0.925
Kruskal-Wallis			p-value = **0.05462**					p-value = **0.3181**					p-value = **0.05431**				

A Tabela 13 apresenta resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 60%, em conjunto com o teste de *Kruskal Wallis*. A partir destes resultados, observa-se que: (**1)** Não existe diferença estatística nos resultados de *Recall*, *Precision* e *F1 Score* dentre as medidas de similaridade.

Tabela 13 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 60%, *log* de eventos sintético e percentual de ruído 30%.

Medida de Similaridade		n	Recall					Precision					FScore				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	2	**0.707**	0.217	0.553	0.707	0.860	**0.837**	0.230	0.675	**0.837**	1.000	**0.766**	0.224	0.608	0.766	0.925
M2	Comparação da dependência de grafos [9]	3	0.666	0.169	0.555	0.584	0.860	0.783	0.188	0.675	**0.675**	1.000	0.720	0.178	0.609	0.626	0.925
M3	Estimativa de características [45]	2	**0.707**	0.217	0.553	0.707	0.860	**0.837**	0.230	0.675	**0.837**	1.000	**0.766**	0.224	0.608	0.766	0.925
M4	Distância da edição de grafos e similaridade de rótulos [48]	1	0.282	NA	0.282	0.282	0.282	0.625	NA	0.625	**0.625**	0.625	0.388	NA	0.388	0.388	0.388
M5	Relação de transições adjacentes [43]	127	0.634	0.147	0.228	0.620	1.000	0.679	0.189	0.029	0.675	1.000	0.637	0.152	0.055	0.639	1.000
Kruskal-Wallis			p-value = **0.5312**					p-value = **0.4545**					p-value = **0.3808**				

A Tabela 14 e Tabela 15 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 80%, em conjunto com os testes de *Kruskal Wallis* e teste *post-hoc* de *Dunn-Bonferroni* respectivamente. A partir destes resultados, observa-se que: **(1)** Não há diferença estatística nos resultados obtidos com *Recall* e *Precision* dentre as medidas de similaridade (cf. Tabela 14). **(2)** M5 possui o maior valor de *Recall*, contudo somente se diferencia estatisticamente de M4 (cf. Tabela 15).

Tabela 14 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 80%, *log* de eventos sintético e percentual de ruído 30%.

Medida de Similaridade		n	Recall					Precision					FScore				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	2	0.707	0.217	0.553	0.707	0.860	**0.837**	0.230	0.675	**0.837**	1.000	**0.766**	0.224	0.608	0.766	0.925
M2	Comparação da dependência de grafos [9]	24	0.710	0.172	0.568	0.606	1.000	0.726	0.178	0.275	0.675	1.000	0.707	0.164	0.424	0.636	1.000
M3	Estimativa de características [45]	2	0.707	0.217	0.553	0.707	0.860	**0.837**	0.230	0.675	**0.837**	1.000	**0.766**	0.224	0.608	0.766	0.925
M4	Distância da edição de grafos e similaridade de rótulos [48]	11	0.593	0.106	0.397	0.579	0.850	0.731	0.117	0.615	0.675	1.000	0.654	0.109	0.482	0.623	0.919
M5	Relação de transições adjacentes [43]	487	**0.733**	0.132	0.395	**0.744**	1.000	0.611	0.196	0.029	0.618	1.000	0.639	0.132	0.055	0.654	1.000
Kruskal-Wallis			p-value = **0.003277**					p-value = **0.0119**					p-value = **0.5853**				

Tabela 15 - Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 80%, *log* de eventos sintético e percentual de ruído 30%.

Comparação	p-value		
	Recall	Precision	Fscore
M1 - M2	0.914764133	0.8333712	0.7376805
M1 - M3	1	1	1
M2 - M3	1	1	0.8430634
M1 - M4	0.485619136	0.7204319	0.8389936
M2 - M4	0.219804265	0.9491699	0.8969072
M3 - M4	0.60702392	0.8233508	1
M1 - M5	1	0.3917891	1
M2 - M5	0.264488905	0.1530319	1
M3 - M5	1	0.5223854	1
M4 - M5	**0.003075247**	0.2988568	1

5.2.2 Síntese dos resultados

O processo de agrupamento de instâncias de processos foi aplicado a *log* de eventos sintético que caracteriza um processo não estruturado, com cerca de 30% de ruído nas instâncias de processos em relação a base de dados anterior. No processo de agrupamento com limiar de similaridade 40%, observou-se que a utilização de medidas de similaridade que consideram os rótulos das atividades (M1 e M3), estrutura dos modelos (M4) não geraram agrupamentos de instâncias de processos, permanecendo o *log* de eventos. Assim como nos resultados apresentados na seção anterior, esse comportamento se explica devido à natureza do *log* de eventos, onde grande parte dos casos continham as mesmas atividades, mas com repetições de

atividades no mesmo caso. Para este limiar, M5 apresentou o maior número de agrupamentos.

No agrupamento com limiar de similaridade 60%, observou-se que o número de agrupamentos gerados foi maior quando comparado ao limiar 40%. M5 se mostrou sensível aos ruídos gerados. Ela gerou um número consideravelmente maior de agrupamentos (cf. Tabela 13), quando comparado ao mesmo processo de agrupamento e percentual de ruído 0% (cf. Tabela 8).

Por fim, a partir dos resultados obtidos, não é possível verificar qual medida obteve o melhor desempenho no agrupamento de instâncias de processos, visto que não houve diferença estatística dos resultados dentre as medidas de similaridade

5.3 *Log* de eventos de hospital universitário

Esta seção irá apresentar os resultados do experimento aplicado no log de eventos reais que representam o processo de um hospital universitário, cujos os registros são de tratamentos de pacientes diagnosticados com câncer. Tal *log* de eventos é composto por 1143 casos, 150.921 eventos, 624 atividades e 981 diferentes fluxos no processo, representando um processo não estruturado.

5.3.1 Análise estatística

A Tabela 16 e Tabela 17 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 40%, em conjunto com os testes de *Kruskal Wallis* e teste post-hoc de *Dunn-Bonferroni* respectivamente. A partir destes resultados, observa-se que: **(1)** há diferença estatística da qualidade média de *Recall, Precision e F1 Score* dos modelos dentre as medidas de similaridade utilizadas no agrupamento de instâncias de processos (cf. Tabela 16). **(2)** M2 e M5 obtiveram *Recall* estatisticamente superiores em relação as demais medidas de similaridade (cf. Tabela 17). **(3)** M5 possui *Precision* superior em relação as demais medidas, contudo a média não se diferencia estatisticamente de M4. **(4)** M5 obteve *F1 Score* estatisticamente superior em relação as demais medidas de similaridade, exceto M4.

(5) As medidas que consideram a relação de dependência de grafos (M2) e a relação dentre as transições adjacentes (M5) resultaram em um número maior de agrupamentos gerados. Esses valores são explicados pela característica do log de eventos que representa um processo não estruturado com muitas atividades e transições entre atividades.

Tabela 16- Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 40%, *log* de eventos reais de um hospital universitário

Medida de Similaridade		n	Recall					Precision					FScore				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	16	0.303	0.316	0.000	0.190	1.000	0.244	0.311	0.000	0.111	1.000	0.239	0.302	0.000	0.134	1.000
M2	Comparação da dependência de grafos [9]	291	**0.854**	0.261	0.013	**0.969**	1.000	0.306	0.238	0.000	0.298	1.000	0.408	0.256	0.000	0.454	1.000
M3	Estimativa de características [45]	9	0.274	0.383	0.001	0.015	1.000	0.176	0.339	0.000	0.007	1.000	0.177	0.333	0.000	0.005	1.000
M4	Distância da edição de grafos e similaridade de rótulos [48]	3	0.197	0.322	0.001	0.022	0.569	0.251	0.431	0.000	0.004	0.749	0.216	0.373	0.000	0.001	0.647
M5	Relação de transições adjacentes [43]	653	**0.854**	0.210	0.032	**0.950**	1.000	**0.408**	0.269	0.000	**0.416**	1.000	**0.502**	0.276	0.000	0.556	1.000
Kruskal-Wallis			p-value = **4.523e-13**					p-value = **6.506e-10**					p-value = **2.948e-11**				

Tabela 17 - Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 40%, *log* de eventos reais de um hospital universitário.

Comparação	p-value		
	Recall	Precision	Fscore
M1 - M2	**3,05317E-08**	0,3362824	**0,04766598**
M1 - M3	0,8231528	0,5566868	0,6979476
M2 - M3	**8,0368E-05**	0,1065298	**0,03722304**
M1 - M4	0,7550131	0,9643428	0,9349931
M2 - M4	**0,003552779**	0,6745915	0,3949121
M3 - M4	0,7455172	0,6892156	0,8585712
M1 - M5	**4,43954E-06**	**0,0120729**	**0,000602628**
M2 - M5	**0,000192324**	**6,85655E-08**	**2,52597E-07**
M3 - M5	**0,001215129**	**0,005542667**	**0,000960934**
M4 - M5	**0,01337988**	0,3731302	0,1278638

A Tabela 18 e Tabela 19 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 60%, em conjunto com os testes de *Kruskal Wallis* e teste *post-hoc* de *Dunn-Bonferroni* respectivamente. A partir destes resultados, observa-se que: **(1)** a qualidade dos modelos foi superior quando comparado ao processo de agrupamento com limiar de similaridade 40%. **(2)** Há diferença estatística de *Recall, Precision e F1 Score* dentre as medidas de

similaridade utilizadas no agrupamento de instâncias de processos (cf. Tabela 18). **(3)** M2 e M5 possuem *Recall* estatisticamente superior em relação as demais medidas (cf. Tabela 19). **(4)** M5 apresenta *Precision* estatisticamente superior em relação as demais medidas. **(5)** M5 possui *F1 Score* estatisticamente superior relacionada as demais medidas, mostrando melhor desempenho para este caso. **(6)** com limiar de similaridade maior 40%, o número de agrupamentos gerados foi consideravelmente maior em todas as medidas de similaridade.

Tabela 18 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 60%, *log* de eventos reais de um hospital universitário.

Medida de Similaridade		n	Recall					Precision					Fscore				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	73	0.502	0.345	0.000	0.481	1.000	0.284	0.323	0.000	**0.109**	1.000	0.306	0.316	0.000	0.177	1.000
M2	Comparação da dependência de grafos [9]	577	**0.860**	0.221	0.024	0.959	1.000	**0.395**	0.250	0.000	**0.387**	1.000	**0.497**	0.259	0.000	0.543	1.000
M3	Estimativa de características [45]	22	0.239	0.312	0.001	0.051	1.000	0.180	0.273	0.000	**0.033**	1.000	0.182	0.271	0.000	0.020	1.000
M4	Distância da edição de grafos e similaridade de rótulos [48]	9	0.084	0.183	0.000	0.019	0.569	0.088	0.248	0.000	**0.005**	0.749	0.088	0.226	0.000	0.009	0.647
M5	Relação de transições adjacentes [43]	837	**0.862**	0.190	0.032	0.947	1.000	**0.446**	0.277	0.000	0.455	1.000	**0.538**	0.277	0.000	0.599	1.000
Kruskal-Wallis			p-value = **2.2e-16**					p-value = **6.962e-14**					p-value = **2.2e-16**				

Tabela 19 - Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 60%, *log* de eventos reais de um hospital universitário.

Comparação	p-value		
	Recall	Precision	Fscore
M1 - M2	**2.50022E-18**	**0.000616208**	**7.01945E-06**
M1 - M3	0.1057328	0.1083853	0.06855212
M2 - M3	**8.01336E-12**	**0.000274204**	**4.5814E-06**
M1 - M4	0.0649285	**0.03610522**	0.06589303
M2 - M4	**1.59133E-07**	**0.000530899**	**0.000515118**
M3 - M4	0.4779589	0.351078	0.5686263
M1 - M5	**9.33594E-15**	**2.18045E-06**	**2.38863E-09**
M2 - M5	**0.009719702**	**0.000615228**	**0.000590817**
M3 - M5	**4.69153E-10**	**8.15553E-06**	**4.70224E-08**
M4 - M5	**1.27564E-06**	**9.25585E-05**	**6.55883E-05**

A Tabela 20 e Tabela 21 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 80%, em conjunto com os testes de *Kruskal Wallis* e teste *post-hoc* de *Dunn-Bonferroni* respectivamente. A partir destes resultados, observa-se que: **(1)** há diferença estatística de *Precision* e *F1 Score* dentre as medidas de similaridade utilizadas no agrupamento de instâncias de

processos (cf. Tabela 20). **(3)** M2 e M5 apresentam *Precision* estatisticamente superior em relação as demais medidas (cf. Tabela 21). **(4)** M2 e M5 possuem valores estatisticamente superior de *F1 Score* em relação as demais medidas, mostrando melhor desempenho para este caso (cf. Tabela 21). **(5)** Com limiar de similaridade maior que 60%, o número de agrupamentos gerados foi consideravelmente maior nas medidas M1, M2, M4 (cf. Tabela 20).

Tabela 20 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 80%, *log* de eventos reais de um hospital universitário.

Medida de Similaridade		n	Recall					Precision					FScore				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	442	0.806	0.254	0.009	0.938	1.000	0.369	0.279	0.000	**0.352**	1.000	**0.449**	0.283	0.000	0.491	1.000
M2	Comparação da dependência de grafos [9]	802	**0.860**	0.194	0.032	0.948	1.000	**0.444**	0.269	0.000	**0.459**	1.000	**0.538**	0.272	0.000	0.605	1.000
M3	Estimativa de características [45]	87	0.483	0.398	0.007	0.414	1.000	0.154	0.264	0.000	**0.033**	1.000	0.170	0.250	0.000	0.053	1.000
M4	Distância da edição de grafos e similaridade de rótulos [48]	664	**0.856**	0.207	0.032	0.950	1.000	**0.411**	0.265	0.000	**0.409**	1.000	**0.505**	0.270	0.000	0.549	1.000
M5	Relação de transições adjacentes [43]	906	**0.864**	0.186	0.032	0.945	1.000	**0.464**	0.287	0.000	0.468	1.000	**0.553**	0.283	0.000	0.612	1.000
Kruskal-Wallis			p-value = **2.775e-12**					p-value = **2.2e-16**					p-value = **2.2e-16**				

Tabela 21 - Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 80%, *log* de eventos reais de um hospital universitário.

Comparação	p-value		
	Recall	Precision	Fscore
M1 - M2	0.0651322	**1.32574E-06**	**2.8606E-08**
M1 - M3	**7.54666E-09**	**5.61611E-11**	**5.86284E-13**
M2 - M3	**2.15765E-12**	**7.05551E-21**	**1.98699E-25**
M1 - M4	**0.03368768**	**0.009489848**	**0.001821796**
M2 - M4	0.7118425	**0.01587214**	**0.009159428**
M3 - M4	**1.49606E-12**	**2.97556E-16**	**7.38636E-20**
M1 - M5	0.102354	**3.14856E-09**	**5.01708E-11**
M2 - M5	0.7147541	0.2331975	0.2730933
M3 - M5	**3.38337E-12**	**4.9747E-23**	**1.16845E-27**
M4 - M5	0.5108159	**0.000377951**	**0.000249076**

5.3.2 Síntese dos resultados

O processo de agrupamento de instâncias de processos foi aplicado no *log* de eventos reais extraídos de processos que continham registros de tratamento de câncer em um hospital. O processo resultante foi caracterizado como um processo não estruturado. Tal processo é composto por muitas atividades e muitos fluxos diferentes. No

processo de agrupamento com limiar de similaridade 40%, observou-se que a utilização de medidas de similaridade que consideram as relações diretas e indiretas dentre as atividades (M2) e a relação de transições adjacentes (5) obtiveram *Recall* estatisticamente superiores em relação as demais medidas. Para *Precision* e *F1 Score* o mesmo comportamento foi observado, contudo comparado a M4 não houve diferença estatística nos resultados.

No agrupamento com limiar de similaridade 60%, observou-se que M2 e M5 obtiveram os melhores resultados comparados as outras medidas, onde M5 se mostra estatisticamente superior nas três métricas de qualidade: *Recall, Precision* e *F1 Score*. Para o processo de agrupamento com limiar de 80%, M2 e M5 mostraram novamente melhores resultados, contudo somente no *Precision* e *F1 Score* se diferenciaram estatisticamente das demais, na qual dentre si, são estatisticamente equivalentes. Por fim, a partir dos resultados obtidos, verifica-se, neste contexto, que M2 e M5 obtiveram melhor desempenho no agrupamento de instâncias de processos.

5.4 *Log* de eventos de processo de faturamento de hospital

Esta seção irá apresentar os resultados do experimento aplicado no log de eventos reais que representa o processo de faturamento de um hospital. Ele contém eventos relacionados ao faturamento de serviços médicos e é composto por 100.000 casos, 451.359 eventos e 18 atividades, representando um processo não estruturado. Os casos são amostras aleatórias de instâncias de processos registradas ao longo de três anos.

5.4.1 Análise estatística

A Tabela 22 e Tabela 23 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 40%, em conjunto com os testes de *Kruskal Wallis* e teste *post-hoc* de *Dunn-Bonferroni* respectivamente. A partir destes resultados, observa-se que: **(1)** Apesar do teste de *Kruskal Wallis* apontar que há diferença estatística da qualidade média de *Recall* e *Precision* dos modelos de

processos (cf. Tabela 22) dentre as medidas de similaridade utilizadas no agrupamento de instâncias de processos, o teste *post-hoc* de *Dunn-Bonferroni* não encontrou diferença estatística na comparação par a par (cf. Tabela 23).

Tabela 22 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 40%, *log* de eventos reais de processo de faturamento de hospital.

Medida de Similaridade		n	Recall					Precision					F1 Score				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	2	0.155	0.021	0.140	0.155	0.169	0.070	0.086	0.010	0.070	0.131	0.077	0.083	0.018	0.077	0.135
M2	Comparação da dependência de grafos [9]	11	0.392	0.233	0.157	0.309	1.000	0.244	0.274	0.000	0.177	1.000	0.271	0.271	0.000	0.201	1.000
M3	Estimativa de características [45]	2	0.274	0.148	0.169	0.274	0.379	0.014	0.016	0.003	0.014	0.026	0.027	0.030	0.005	0.027	0.048
M4	Distância da edição de grafos e similaridade de rótulos [48]	1	0.129	NA	0.129	0.129	0.129	0.002	NA	0.002	0.002	0.002	0.004	NA	0.004	0.004	0.004
M5	Relação de transições adjacentes [43]	91	**0.580**	0.241	0.107	0.558	1.000	**0.443**	0.275	0.000	0.392	1.000	**0.488**	0.260	0.000	0.460	1.000
Kruskal-Wallis			p-value = **0.00196**					p-value = **0.0005259**					p-value = **0.0003302**				

Tabela 23 - Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 40%, *log* de eventos reais de processo de faturamento de hospital.

Comparação	p-value		
	Recall	Precision	F1 Score
M1 - M2	0.45206938	0.5530674	0.5626194
M1 - M3	0.77156187	1	1
M2 - M3	0.8216115	0.52780506	0.57480923
M1 - M4	0.93705119	1	1
M2 - M4	0.54217655	0.52632622	0.55673684
M3 - M4	0.69582471	0.94752613	0.94752613
M1 - M5	0.07602681	0.07799715	0.06782302
M2 - M5	0.11716559	0.06154753	**0.03905064**
M3 - M5	0.20982052	0.07660131	0.07494414
M4 - M5	0.22819884	0.17808079	0.17588234

A Tabela 24 e Tabela 25 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 60%, em conjunto com os testes de *Kruskal Wallis* e teste *post-hoc* de *Dunn-Bonferroni* respectivamente. A partir destes resultados, observa-se que: **(1)** Não há diferença estatística da qualidade média de *Recall*, *Precision* e F1 Score (cf. Tabela 24) dentre as medidas de similaridade utilizadas no agrupamento de instâncias de processos.

Tabela 24 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 60%, *log* de eventos reais de processo de faturamento de hospital.

Medida de Similaridade		n	Recall					Precision					F1 Score				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	10	0.434	0.275	0.149	0.353	0.853	0.413	0.343	0.017	0.321	1.000	0.404	0.314	0.030	0.370	0.900
M2	Comparação da dependência de grafos [9]	35	0.494	0.222	0.146	0.444	1.000	0.368	0.295	0.000	0.290	1.000	0.401	0.272	0.000	0.351	1.000
M3	Estimativa de características [45]	8	0.553	0.244	0.135	0.559	0.853	0.448	0.373	0.002	0.375	1.000	0.460	0.342	0.004	0.421	0.900
M4	Distância da edição de grafos e similaridade de rótulos [48]	4	0.201	0.064	0.119	0.209	0.268	0.067	0.060	0.002	0.059	0.146	0.091	0.077	0.003	0.086	0.189
M5	Relação de transições adjacentes [43]	267	**0.681**	0.200	0.195	0.677	1.000	**0.527**	0.242	0.000	0.456	1.000	**0.584**	0.220	0.000	0.543	1.000
Kruskal-Wallis			p-value = **3.706e-08**					p-value = **7.988e-06**					p-value = **5.959e-07**				

Tabela 25 - Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 60%, *log* de eventos reais de processo de faturamento de hospital.

Comparação	p-value		
	Recall	Precision	F1 Score
M1 - M2	0.7439818	0.579106568	0.831017937
M1 - M3	0.4399674	0.750359485	0.595937995
M2 - M3	0.4530951	0.472870026	0.431641152
M1 - M4	0.2469334	0.08270237	0.123890598
M2 - M4	0.1307291	0.101656331	0.127100358
M3 - M4	0.08543827	0.070501583	0.068984872
M1 - M5	**0.01112833**	0.211281853	0.089952065
M2 - M5	**4.08937E-05**	**0.0005933**	**0.0001142**
M3 - M5	0.2290429	0.427570656	0.346890024
M4 - M5	**0.001752197**	**0.0025142**	**0.0022398**

A Tabela 26 e Tabela 27 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 80%, em conjunto com os testes de *Kruskal Wallis* e teste *post-hoc* de *Dunn-Bonferroni* respectivamente. A partir destes resultados, observa-se que: **(1)** Há diferença estatística da qualidade média de *Recall*, *Precision* e F1 Score dos modelos (cf. Tabela 26) dentre as medidas de similaridade utilizadas no agrupamento de instâncias de processos. **(2)** M5 apresentou superioridade estatística em relação as demais medidas nas três métricas de qualidade: *Recall*, *Precision* e *F1 Score*.

Tabela 26 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 80%, *log* de eventos reais de processo de faturamento de hospital.

Medida de Similaridade		n	Recall					Precision					F1 Score				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	34	0.503	0.217	0.143	0.482	1.000	0.425	0.294	0.000	0.368	1.000	0.435	0.267	0.000	0.387	1.000
M2	Comparação da dependência de grafos [9]	133	0.622	0.221	0.170	0.614	1.000	0.498	0.277	0.000	0.416	1.000	0.541	0.255	0.000	0.503	1.000
M3	Estimativa de características [45]	21	0.450	0.223	0.140	0.442	0.853	0.378	0.296	0.000	0.307	1.000	0.374	0.263	0.000	0.323	0.900
M4	Distância da edição de grafos e similaridade de rótulos [48]	67	0.570	0.251	0.121	0.527	1.000	0.468	0.296	0.000	0.361	1.000	0.501	0.280	0.000	0.427	1.000
M5	Relação de transições adjacentes [43]	539	**0.753**	0.184	0.228	0.792	1.000	**0.563**	0.222	0.000	0.508	1.000	**0.633**	0.199	0.000	0.612	1.000
Kruskal-Wallis			p-value = **2.2e-16**					p-value = **3.172e-08**					p-value = **3.585e-13**				

Tabela 27 - Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 80%, *log* de eventos reais de processo de faturamento de hospital.

Comparação	p-value		
	Recall	Precision	F1 Score
M1 - M2	0.02436787	0.293876358	0.09236191
M1 - M3	0.5966499	0.531999402	0.4740759
M2 - M3	**0.0173921**	0.103857564	**0.03156913**
M1 - M4	0.1846979	0.576119455	0.3714473
M2 - M4	0.3026689	0.428460523	0.3569321
M3 - M4	0.1013005	0.298378616	0.1349694
M1 - M5	**9.47164E-09**	**0.0015786**	**5.64047E-05**
M2 - M5	**9.13904E-09**	**0.0012452**	**1.79454E-05**
M3 - M5	**1.36628E-07**	**0.0015865**	**3.36605E-05**
M4 - M5	**1.67551E-08**	**0.0017228**	**2.24612E-05**

5.4.2 Síntese dos resultados

O processo de agrupamento de instâncias de processos foi aplicado no *log* de eventos reais extraídos do processo de faturamento de um hospital. O processo resultante é caracterizado como um processo não estruturado. Ele é composto por poucas atividades e muitos fluxos diferentes. Somente houve diferença estatística dos resultados no limiar de similaridade de 80%, onde M5 obteve melhor desempenho. Desta forma, não é possível verificar qual medida obteve o melhor desempenho no agrupamento de instâncias de processos

5.5 *Log* de eventos de processo de pedido de recebimento

Esta seção irá apresentar os resultados do experimento aplicado no log de eventos reais que representa o processo de pedido de recebimento de construção em um município. Ele é composto por 1434 casos, 8577 eventos, 27 atividades e 116 diferentes fluxos no processo, representando um processo semiestruturado.

5.5.1 Análise estatística

A Tabela 28 e Tabela 29 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 40%, em conjunto com os testes de *Kruskal Wallis* e teste *post-hoc* de *Dunn-Bonferroni* respectivamente. A partir destes resultados, observa-se que: **(1)** Apesar do teste de *Kruskal Wallis* apontar que existe diferença estatística da qualidade média de *Recall*, *Precision* e *F1 Score* dos modelos (cf. Tabela 28) dentre as medidas de similaridade utilizadas no agrupamento de instâncias de processos, o teste *post-hoc* de *Dunn-Bonferroni* não mostrou diferença estatística na comparação par a par (cf. Tabela 29).

Tabela 28- Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 40%, *log* de eventos reais de um processo de recebimento.

Medida de Similaridade		n	Recall					Precision					F1 Score				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	3	0.453	0.124	0.331	0.449	0.580	0.171	0.114	0.048	0.192	0.272	0.237	0.135	0.084	0.288	0.339
M2	Comparação da dependência de grafos [9]	3	0.442	0.176	0.267	0.441	0.619	0.222	0.150	0.048	0.307	0.310	0.286	0.178	0.082	0.362	0.413
M3	Estimativa de características [45]	1	0.308	NA	0.308	0.308	0.308	0.049	NA	0.049	0.049	0.049	0.085	NA	0.085	0.085	0.085
M4	Distância da edição de grafos e similaridade de rótulos [48]	1	0.308	NA	0.308	0.308	0.308	0.049	NA	0.049	0.049	0.049	0.085	NA	0.085	0.085	0.085
M5	Relação de transições adjacentes [43]	28	**0.717**	0.199	0.365	0.798	1.000	**0.529**	0.221	0.239	0.448	1.000	**0.596**	0.198	0.308	0.573	1.000
Kruskal-Wallis			p-value = **0.02478**					p-value = **0.005666**					p-value = **0.006474**				

Tabela 29 - Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 40%, *log* de eventos reais de um processo de recebimento.

Comparação	p-value		
	Recall	Precision	F1 Score
M1 - M2	1	1	0.9569886
M1 - M3	0.8958579	1	0.9779976
M2 - M3	0.7178419	0.93022432	1
M1 - M4	1	1	1
M2 - M4	0.8203907	1	1
M3 - M4	1	1	1
M1 - M5	0.3572607	0.07767009	0.0920445
M2 - M5	0.5655068	0.20825723	0.2110695
M3 - M5	0.1894998	0.20954276	0.2110377
M4 - M5	0.2526663	0.27939034	0.2813836

A Tabela 30 e Tabela 31 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 60%, em conjunto com os testes de *Kruskal Wallis* e teste *post-hoc* de *Dunn-Bonferroni* respectivamente. A partir destes resultados, observa-se que: **(1)** Há diferença estatística da qualidade média de *Recall*, *Precision* e *F1 Score* dos modelos (cf. Tabela 30) dentre as medidas de similaridade utilizadas no agrupamento de instâncias de processos. **(2)** M5 obteve valores superiores nas três métricas de qualidade: *Recall*, *Precision* e *F1 Score*, contudo não se diferenciou estatisticamente somente de M3.

Tabela 30 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 60%, *log* de eventos reais de um processo de recebimento.

Medida de Similaridade		n	Recall					Precision					F1 Score				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	4	0.595	0.223	0.333	0.587	0.875	0.285	0.201	0.049	0.290	0.509	0.371	0.235	0.086	0.377	0.644
M2	Comparação da dependência de grafos [9]	11	0.629	0.177	0.269	0.622	0.880	0.445	0.278	0.136	0.378	1.000	0.503	0.221	0.180	0.450	0.907
M3	Estimativa de características [45]	1	0.308	NA	0.308	0.308	0.308	0.049	NA	0.049	0.049	0.049	0.085	NA	0.085	0.085	0.085
M4	Distância da edição de grafos e similaridade de rótulos [48]	2	0.304	0.062	0.260	0.304	0.348	0.150	0.094	0.084	0.150	0.217	0.197	0.099	0.127	0.197	0.267
M5	Relação de transições adjacentes [43]	57	**0.888**	0.117	0.535	0.924	1.000	**0.713**	0.251	0.239	0.630	1.000	**0.777**	0.196	0.342	0.720	1.000
Kruskal-Wallis			p-value = **7.091e-06**					p-value = **6.214e-05**					p-value = **3.043e-05**				

Tabela 31 - Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 60%, *log* de eventos reais de um processo de recebimento.

Comparação	p-value		
	Recall	Precision	F1 Score
M1 - M2	1	0.67600894	0.840605579
M1 - M3	0.742727884	0.76781767	0.751744617
M2 - M3	0.764517077	0.57828734	0.648100597
M1 - M4	0.819274082	0.71496802	0.674605641
M2 - M4	0.798891066	0.5726161	0.644097398
M3 - M4	1	0.8943426	0.909778669
M1 - M5	**0.0272191**	**0.0232125**	**0.0255005**
M2 - M5	**0.0006428**	**0.0150069**	**0.0058297**
M3 - M5	0.128879652	0.1110278	0.109741247
M4 - M5	**0.0316826**	**0.0317568**	**0.0282464**

A Tabela 32 e Tabela 33 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 80%, em conjunto com os testes de *Kruskal Wallis* e teste *post-hoc* de *Dunn-Bonferroni* respectivamente. A partir destes resultados, observa-se que: **(1)** Há diferença estatística da qualidade média de *Recall*, *Precision* e *F1 Score* dos modelos (cf. Tabela 33) dentre as medidas de similaridade utilizadas no agrupamento de instâncias de processos. **(2)** M5 obteve valores estatisticamente superiores nas três métricas de qualidade: *Recall*, *Precision* e *F1 Score*.

Tabela 32 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 80%, *log* de eventos reais de um processo de recebimento.

Medida de Similaridade		n	Recall					Precision					F1 Score				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	12	0.682	0.199	0.334	0.633	1.000	0.478	0.343	0.072	0.321	1.000	0.536	0.293	0.118	0.417	1.000
M2	Comparação da dependência de grafos [9]	33	0.800	0.149	0.443	0.814	1.000	0.585	0.249	0.188	0.509	1.000	0.662	0.206	0.264	0.644	1.000
M3	Estimativa de características [45]	3	0.333	0.042	0.292	0.331	0.376	0.135	0.099	0.053	0.107	0.245	0.182	0.105	0.092	0.156	0.297
M4	Distância da edição de grafos e similaridade de rótulos [48]	26	0.740	0.173	0.377	0.713	1.000	0.557	0.247	0.239	0.524	1.000	0.620	0.211	0.304	0.568	1.000
M5	Relação de transições adjacentes [43]	95	**0.920**	0.100	0.393	0.937	1.000	**0.744**	0.218	0.239	0.710	1.000	**0.812**	0.168	0.322	0.806	1.000
Kruskal-Wallis			p-value = **3.51e-10**					p-value = **1.787e-06**					p-value = **6.182e-08**				

Tabela 33 - Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 80%, *log* de eventos reais de um processo de recebimento.

Comparação	p-value		
	Recall	Precision	F1 Score
M1 - M2	0.2693347	0.384140444	0.399288879
M1 - M3	0.2240991	0.165394889	0.174158685
M2 - M3	0.05451197	**0.048365**	0.053888431
M1 - M4	0.6278674	0.516010633	0.589472743
M2 - M4	0.3816977	0.734845129	0.638088571
M3 - M4	0.1261908	0.061980444	**0.086026**
M1 - M5	**0.000171895**	**0.0036733**	**0.0010835**
M2 - M5	**0.00023338**	**0.0029429**	**0.0011927**
M3 - M5	**0.000598195**	**0.0059017**	**0.001328**
M4 - M5	**1.30206E-05**	**0.0038526**	**0.0005821**

5.5.2 Síntese dos resultados

O processo de agrupamento de instâncias de processos foi aplicado no *log* de eventos reais de um processo semiestruturado de pedido de recebimento de construção em um município. Nos três limiares aplicados no processo de agrupamento, observou-se superioridade da medida de similaridade que consideram as relações de transições adjacentes (M5) nas três métricas de qualidade: *Recall, Precision* e *F1 Score*. No limiar de similaridade de 60%, M5 mostrou superioridade estatística em relação as outras medidas, exceto com M3. No limiar de similaridade de 80%, M5 se mostrou estatisticamente superior a todas as outras medidas. Desta forma, visto a

predominância dos resultados de M5, verifica-se, neste contexto, que a mesma obteve o melhor desempenho no agrupamento de instâncias de processos.

5.6 *Log* de eventos de tratamento de pacientes em Hospital

Esta seção irá apresentar os resultados do experimento aplicado no log de eventos reais que representa o processo de tratamento de pacientes no hospital e contém eventos de casos de infecções. Neste caso cada instância de processo representa um caminho de um paciente no hospital. Tal *log* de eventos é composto por 15214 casos, 1050 eventos, 16 atividades e 846 diferentes fluxos no processo, representando um processo não estruturado.

5.6.1 Análise estatística

A Tabela 34 e Tabela 35 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 40%, em conjunto com os testes de *Kruskal Wallis* e teste *post-hoc* de *Dunn-Bonferroni* respectivamente. A partir destes resultados, observa-se que: **(1)** Apesar do teste de *Kruskal Wallis* apontar que existe diferença estatística da qualidade média de *Recall*, *Precision* e *F1 Score* dos modelos (cf. Tabela 34) dentre as medidas de similaridade utilizadas no agrupamento de instâncias de processos, o teste *post-hoc* de *Dunn-Bonferroni* não mostrou diferença estatística na comparação par a par (cf. Tabela 35).

Tabela 34 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 40%, *log* de eventos reais de processo de tratamento de pacientes.

Medida de Similaridade		n	Recall					Precision					F1 Score				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	1	0.284	NA	0.284	0.284	0.284	0.049	NA	0.049	0.049	0.049	0.084	NA	0.084	0.084	0.084
M2	Comparação da dependência de grafos [9]	3	0.324	0.056	0.286	0.297	0.388	0.166	0.102	0.050	0.207	0.240	0.207	0.105	0.085	0.266	0.270
M3	Estimativa de características [45]	1	0.284	NA	0.284	0.284	0.284	0.049	NA	0.049	0.049	0.049	0.084	NA	0.084	0.084	0.084
M4	Distância da edição de grafos e similaridade de rótulos [48]	1	0.284	NA	0.284	0.284	0.284	0.049	NA	0.049	0.049	0.049	0.084	NA	0.084	0.084	0.084
M5	Relação de transições adjacentes [43]	101	**0.568**	0.183	0.269	0.552	1.000	**0.431**	0.213	0.063	0.376	1.000	**0.480**	0.194	0.103	0.442	1.000
Kruskal-Wallis			p-value = **0.007989**					p-value = **0.004101**					p-value = **0.002875**				

Tabela 35 - Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 40%, *log* de eventos reais de processo de tratamento de pacientes.

Comparação	p-value		
	Recall	Precision	F1 Score
M1 - M2	1	1	1
M1 - M3	1	1	1
M2 - M3	1	1	1
M1 - M4	1	1	1
M2 - M4	1	1	1
M3 - M4	1	1	1
M1 - M5	0.2275499	0.19645145	0.19539348
M2 - M5	0.1622098	0.09961752	0.06369464
M3 - M5	0.3033999	0.26193526	0.26052464
M4 - M5	0.4550998	0.39290289	0.39078697

A Tabela 36 e Tabela 37 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 60%, em conjunto com os testes de *Kruskal Wallis* e teste *post-hoc* de *Dunn-Bonferroni* respectivamente. A partir destes resultados, observa-se que: **(1)** Não há diferença estatística da qualidade do *Recall Precision e F1 Score* (cf. Tabela 36 e Tabela 37) dentre as medidas de similaridade utilizadas no agrupamento de instâncias de processos.

Tabela 36 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 60%, *log* de eventos real de processo de tratamento de pacientes.

Medida de Similaridade		n	Recall					Precision					F1 Score				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	2	0.408	0.183	0.279	0.408	0.537	0.190	0.199	0.049	0.190	0.331	0.246	0.231	0.083	0.246	0.409
M2	Comparação da dependência de grafos [9]	11	0.429	0.143	0.297	0.376	0.833	0.286	0.208	0.063	0.273	0.842	0.329	0.192	0.103	0.337	0.838
M3	Estimativa de características [45]	2	0.328	0.070	0.278	0.328	0.378	0.249	0.283	0.049	0.249	0.450	0.247	0.231	0.083	0.247	0.411
M4	Distância da edição de grafos e similaridade de rótulos [48]	1	0.284	NA	0.284	0.284	0.284	0.049	NA	0.049	0.049	0.049	0.084	NA	0.084	0.084	0.084
M5	Relação de transições adjacentes [43]	335	**0.663**	0.194	0.080	0.645	1.000	**0.538**	0.202	0.063	0.503	1.000	**0.584**	0.191	0.103	0.555	1.000
Kruskal-Wallis			p-value = 2.762e-05					p-value = 1.391e-05					p-value = 9.705e-06				

Tabela 37 - Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 60%, *log* de eventos reais de processo de tratamento de pacientes.

Comparação	p-value		
	Recall	Precision	F1 Score
M1 - M2	0.93589502	0.941157516	1
M1 - M3	0.949932053	0.959477508	0.992135843
M2 - M3	1	0.903235952	1
M1 - M4	1	0.856282635	0.902600223
M2 - M4	1	1	1
M3 - M4	1	1	1
M1 - M5	0.235850259	0.14700376	0.186828994
M2 - M5	**0.00068728**	**0.000212368**	**0.000283186**
M3 - M5	0.12614504	0.28506185	0.128848041
M4 - M5	0.230845681	0.259162746	0.194104453

A Tabela 38 e Tabela 39 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 80%, em conjunto com os testes de *Kruskal Wallis* e teste *post-hoc* de *Dunn-Bonferroni* respectivamente. A partir destes resultados, observa-se que: **(1)** Há diferença estatística da qualidade do *Recall Precision e F1 Score* (cf. Tabela 38) dentre as medidas de similaridade utilizadas no agrupamento de instâncias de processos. **(2)** M5 somente apresenta diferença estatística no *Recall* entre M2 e M4. **(3)** M5 e M3 não se diferenciam estatisticamente no *Precision* e *F1 Score*.

Tabela 38 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 80%, *log* de eventos reais de processo de tratamento de pacientes.

Medida de Similaridade		n	Recall					Precision					F1 Score				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	7	0.546	0.211	0.350	0.484	1.000	0.407	0.308	0.094	0.301	1.000	0.448	0.275	0.148	0.371	1.000
M2	Comparação da dependência de grafos [9]	101	0.597	0.202	0.120	0.568	1.000	0.467	0.229	0.063	0.411	1.000	0.512	0.211	0.103	0.471	1.000
M3	Estimativa de características [45]	3	0.490	0.233	0.296	0.425	0.749	0.347	0.375	0.049	0.224	0.768	0.378	0.345	0.084	0.293	0.759
M4	Distância da edição de grafos e similaridade de rótulos [48]	35	0.536	0.116	0.356	0.500	0.903	0.367	0.143	0.135	0.364	0.954	0.430	0.132	0.200	0.410	0.928
M5	Relação de transições adjacentes [43]	642	**0.724**	0.206	0.080	0.722	1.000	**0.578**	0.219	0.063	0.522	1.000	**0.632**	0.205	0.103	0.593	1.000
Kruskal-Wallis			p-value = **9.59e-14**					p-value = **5.097e-16**					p-value = **3.727e-16**				

Tabela 39 - Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 80%, *log* de eventos reais de processo de tratamento de pacientes.

Comparação	p-value		
	Recall	Precision	F1 Score
M1 - M2	0.7432021	0.8710605	0.7884513
M1 - M3	1	0.9952774	0.9697519
M2 - M3	0.6881176	0.8308548	0.9489916
M1 - M4	0.9658665	0.798885	0.9008924
M2 - M4	0.155017	0.04062936	0.07234073
M3 - M4	0.9388389	0.7697043	0.8579576
M1 - M5	0.05344461	**0.04815653**	**0.04307053**
M2 - M5	**8.11243E-08**	**1.15397E-08**	**2.7246E-09**
M3 - M5	0.1505126	0.2451965	0.224913
M4 - M5	**7.35856E-08**	**7.17341E-10**	**3.06763E-09**

5.6.2 Síntese dos resultados

O processo de agrupamento de instâncias de processos foi aplicado em um *log* de eventos reais de um processo de tratamento de pacientes no hospital, cujos os eventos são de casos de infecções. O processo é semiestruturado.

Nos três limiares de similaridades aplicados no processo de agrupamento, não se observou medidas de similaridade com melhor desempenho superior estatisticamente as demais. Desta forma, não é possível verificar qual medida obteve o melhor desempenho no agrupamento de instâncias de processos

5.7 *Log* de eventos de processo de pedido de empréstimo

Esta seção irá apresentar os resultados do experimento aplicado no log de eventos reais que representa o processo de pedido de empréstimo de uma instituição financeira. As instâncias de processo representam os clientes que solicitam empréstimos, e as atividades descrevem as etapas realizadas para os clientes. Tal *log* de eventos é composto por 13087 casos, 262200 eventos, 37 atividades e 4366 diferentes fluxos no processo, representando um processo não estruturado.

5.7.1 Análise estatística

A Tabela 40 apresenta os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 40%, em conjunto com o teste de *Kruskal Wallis*, observa-se que: **(1)** Não há diferença estatística da qualidade média de *Recall, Precision e F1 Score* dos modelos (cf. Tabela 40) dentre as medidas de similaridade utilizadas no agrupamento de instâncias de processos.

Tabela 40 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 40%, *log* de eventos reais de processo de pedido de empréstimo.

Medida de Similaridade		n	Recall					Precision					FScore				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	2	**0.373**	0.162	0.259	0.373	0.487	**0.364**	0.020	0.350	0.364	0.379	0.362	0.091	0.297	**0.362**	0.426
M2	Comparação da dependência de grafos [9]	3	0.347	0.093	0.276	0.312	0.452	0.241	0.136	0.146	0.179	0.396	0.262	0.075	0.217	0.221	0.349
M3	Estimativa de características [45]	2	0.241	0.058	0.200	0.241	0.283	0.291	0.148	0.186	0.291	0.395	0.245	0.029	0.224	0.245	0.266
M4	Distância da edição de grafos e similaridade de rótulos [48]	1	0.349	NA	0.349	0.349	0.349	0.197	NA	0.197	0.197	0.197	0.252	NA	0.252	0.252	0.252
M5	Relação de transições adjacentes [43]	19	0.366	0.191	0.147	0.339	1.000	**0.472**	0.174	0.304	0.375	1.000	**0.405**	0.179	0.208	**0.344**	1.000
Kruskal-Wallis			p-value = 0.7075					p-value = **0.1918**					p-value = 0.1203				

A Tabela 41 e Tabela 42 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 60%, em conjunto com os testes de *Kruskal Wallis* e teste *post-hoc* de *Dunn-Bonferroni* respectivamente. A partir destes resultados, observa-se que: **(1)** Não há diferença estatística da qualidade do *Recall* e *F1 Score* dos modelos (cf. Tabela 41) dentre as medidas de similaridade utilizadas no agrupamento de instâncias de processos. **(2)** Apesar do teste de *Kruskal Wallis* apontar para existência de diferença estatística no resultado de *Precision* (cf. Tabela 41), o teste *post-hoc* de *Dunn-Bonferroni* não mostrou diferenças na comparação par a par (cf. Tabela 42).

Tabela 41 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos com limiar de similaridade de 60% para *log* de eventos real de processo de pedido de empréstimo

Medida de Similaridade		n	Recall					Precision					FScore				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	3	**0.388**	0.130	0.309	0.316	0.538	0.495	0.204	0.355	0.401	0.729	0.434	0.161	0.334	0.349	0.620
M2	Comparação da dependência de grafos [9]	8	0.376	0.137	0.238	0.338	0.596	0.449	0.132	0.307	0.409	0.730	0.406	0.131	0.286	0.367	0.656
M3	Estimativa de características [45]	2	0.241	0.058	0.200	0.241	0.283	0.291	0.148	0.186	0.291	0.395	0.245	0.029	0.224	0.245	0.266
M4	Distância da edição de grafos e similaridade de rótulos [48]	2	0.274	0.039	0.247	0.274	0.301	0.213	0.032	0.190	0.213	0.235	0.240	0.035	0.215	0.240	0.264
M5	Relação de transições adjacentes [43]	53	**0.360**	0.162	0.128	**0.369**	1.000	**0.541**	0.172	0.302	**0.549**	1.000	**0.421**	0.155	0.184	**0.431**	1.000
Kruskal-Wallis			p-value = 0.5923					p-value = **0.03865**					p-value = 0.113				

Tabela 42 - Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 60%, *log* de eventos reais de processo de pedido de empréstimo.

Comparação	p-value
	Recall
M1 - M2	0.7518037
M1 - M3	0.3954446
M2 - M3	0.4481238
M1 - M4	0.2920364
M2 - M4	0.265655
M3 - M4	0.7886471
M1 - M5	0.7135285
M2 - M5	0.374445
M3 - M5	0.3591256
M4 - M5	0.1361759

A Tabela 43 e Tabela 44 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 80%, em conjunto com os testes de *Kruskal Wallis* e teste *post-hoc* de *Dunn-Bonferroni* respectivamente. A partir destes resultados, observa-se que: **(1)** Não há diferença estatística da qualidade média de *Recall, Precision* e *F1 Score* dos modelos (cf. Tabela 43) dentre as medidas de similaridade utilizadas no agrupamento de instâncias de processos.

Tabela 43 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 80%, *log* de eventos real de processo de pedido de empréstimo.

Medida de Similaridade		n	Recall					Precision					FScore				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	9	**0.415**	0.123	0.231	0.444	0.655	**0.571**	0.128	0.409	**0.536**	0.765	**0.477**	0.118	0.296	**0.469**	0.662
M2	Comparação da dependência de grafos [9]	26	0.381	0.123	0.169	0.358	0.615	**0.555**	0.099	0.414	**0.551**	0.715	0.446	0.110	0.244	0.419	0.632
M3	Estimativa de características [45]	5	0.279	0.085	0.204	0.268	0.423	0.279	0.106	0.177	0.256	0.404	0.259	0.037	0.218	0.250	0.315
M4	Distância da edição de grafos e similaridade de rótulos [48]	16	0.401	0.151	0.154	0.366	0.659	0.495	0.129	0.303	0.505	0.715	0.435	0.133	0.213	0.405	0.623
M5	Relação de transições adjacentes [43]	133	0.397	0.146	0.091	0.403	1.000	**0.574**	0.149	0.301	**0.540**	1.000	0.457	0.136	0.152	0.456	1.000
Kruskal-Wallis			p-value = 0.3569					p-value = **0.002854**					p-value = 0.02032				

Tabela 44 - Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 80%, *log* de eventos reais de processo de pedido de empréstimo.

Comparação	p-value	
	Precision	Fscore
M1 - M2	1	0.884768036
M1 - M3	**0.0112857**	**0.0127334**
M2 - M3	**0.0054555**	**0.0105725**
M1 - M4	0.323999926	0.893591903
M2 - M4	0.240414693	0.81434175
M3 - M4	0.068242868	**0.0191555**
M1 - M5	0.976848205	0.776622557
M2 - M5	1	0.818075254
M3 - M5	**0.0030629**	**0.0089549**
M4 - M5	0.104527997	0.827078503

5.7.2 Considerações finais

O processo de agrupamento de instâncias de processos foi aplicado em um *log* de eventos reais de um processo de pedido de empréstimo de uma instituição financeira. O processo é caracterizado como um processo semiestruturado. Nos limiares de similaridade de 40%, 60% e 80% aplicados no processo de agrupamento, não se observou superioridade estatística dos resultados dentre as medidas de similaridade. Desta forma, não é possível verificar estatisticamente qual medida de similaridade obteve o melhor desempenho no agrupamento de instâncias de processos.

5.8 *Log* de eventos de multas de tráfego rodoviário

Esta seção irá apresentar os resultados do experimento aplicado no log de eventos reais extraídos de um processo de gerenciamento de multas de tráfego rodoviário. Cada instância de processo representa uma multa, que pode ser contestada, e deve ser eventualmente paga. Tal *log* de eventos é composto por 150370 casos, 561470 eventos, 11 atividades e 231 diferentes fluxos no processo, representando um processo semiestruturado.

5.8.1 Análise estatística

A Tabela 45 e Tabela 46 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 40%, em conjunto com os testes de *Kruskal Wallis* e teste *post-hoc* de *Dunn-Bonferroni* respectivamente. A partir destes resultados, observa-se que: **(1)** O número de agrupamentos gerados por M5 é expressivamente maior que as demais medidas; **(2)** Não há diferença estatística da qualidade média de *Recall*, *Precision* e *F1 Score* dos modelos (cf. Tabela 45) dentre as medidas de similaridade utilizadas no agrupamento de instâncias de processos.

Tabela 45 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 40%, *log* de eventos reais de multas de tráfego rodoviário.

Medida de Similaridade		n	Recall					Precision					FScore				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	2	0.380	0.001	0.380	0.380	0.381	0.395	0.120	0.310	0.395	0.480	0.383	0.058	0.342	0.383	0.424
M2	Comparação da dependência de grafos [9]	3	0.388	0.039	0.345	0.401	0.419	0.483	0.006	0.479	0.480	0.490	0.430	0.025	0.401	0.441	0.447
M3	Estimativa de características [45]	2	0.380	0.007	0.375	0.380	0.385	0.389	0.129	0.297	0.389	0.480	0.378	0.061	0.336	0.378	0.421
M4	Distância da edição de grafos e similaridade de rótulos [48]	1	0.406	NA	0.406	0.406	0.406	0.174	NA	0.174	0.174	0.174	0.243	NA	0.243	0.243	0.243
M5	Relação de transições adjacentes [43]	29	**0.783**	0.145	0.470	0.782	1.000	**0.877**	0.135	0.538	**0.938**	1.000	0.821	0.127	0.574	0.815	1.000
Kruskal-Wallis			p-value = 0.001016					p-value = **0.0008751**					p-value = 0.0009651				

Tabela 46- Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 40%, *log* de eventos reais de multas de tráfego rodoviário.

Comparação	p-value		
	Recall	Precision	Fscore
M1 - M2	0.97696825	0.97605053	0.94746857
M1 - M3	1	0.96288129	0.92634921
M2 - M3	1	1	1
M1 - M4	1	1	1
M2 - M4	1	1	1
M3 - M4	1	1	1
M1 - M5	**0.04559419**	0.0616941	0.06442341
M2 - M5	0.06077904	**0.0908087**	0.1107502
M3 - M5	0.06839128	0.07782753	0.06839128
M4 - M5	0.36475098	0.11021724	0.11389172

A Tabela 47 e Tabela 48 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 60%, em conjunto com os testes de *Kruskal Wallis* e teste *post-hoc* de *Dunn-Bonferroni* respectivamente. A partir destes resultados, observa-se que: **(1)** Há diferença estatística da qualidade média de *Recall*, *Precision* e *F1 Score* dos modelos (cf. Tabela 47) entre as medidas de similaridade utilizadas no agrupamento de instâncias de processos; **(2)** M5 resultou em um *Recall* estatisticamente superior as demais medidas; **(3)** M5 apresenta *Precision* estatisticamente superior as demais medidas, exceto M1; **(4)** M5 resultou em *F1 Score* estatisticamente superior as demais medidas de similaridade.

Tabela 47 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 60%, *log* de eventos reais de multas de tráfego rodoviário.

Medida de Similaridade		n	Recall					Precision					FScore				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	5	0.600	0.135	0.493	0.514	0.800	0.822	0.216	0.471	0.913	1.000	0.678	0.127	0.492	0.670	0.810
M2	Comparação da dependência de grafos [9]	8	0.682	0.139	0.427	0.685	0.893	0.771	0.257	0.399	0.884	1.000	0.710	0.176	0.412	0.744	0.943
M3	Estimativa de características [45]	3	0.541	0.227	0.379	0.444	0.800	0.726	0.231	0.480	0.760	0.938	0.602	0.178	0.424	0.602	0.779
M4	Distância da edição de grafos e similaridade de rótulos [48]	2	0.336	0.070	0.286	0.336	0.385	0.457	0.032	0.435	0.457	0.480	0.386	0.058	0.345	0.386	0.427
M5	Relação de transições adjacentes [43]	74	**0.904**	0.108	0.585	0.936	1.000	**0.936**	0.123	0.320	1.000	1.000	**0.915**	0.110	0.457	0.953	1.000
Kruskal-Wallis			p-value = 7.352e-07					p-value = **0.003493**					p-value = 1.782e-06				

Tabela 48 - Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 60%, *log* de eventos reais de multas de tráfego rodoviário.

Comparação	p-value		
	Recall	Precision	Fscore
M1 - M2	0.776075777	0.9048065	0.891575871
M1 - M3	0.947681702	0.6059914	0.819694962
M2 - M3	0.864214335	0.5580442	0.826673161
M1 - M4	1	0.4093246	0.958330542
M2 - M4	0.835266226	0.3484733	0.781355609
M3 - M4	0.976604199	0.6676283	0.822480779
M1 - M5	**0.0036656**	0.3003506	**0.0066461**
M2 - M5	**0.0031919**	**0.115394**	**0.0062163**
M3 - M5	**0.015793**	**0.135391**	**0.0062163**
M4 - M5	**0.0190622**	**0.130261**	**0.0062163**

A Tabela 49 e Tabela 50 apresentam os resultados obtidos no agrupamento de instâncias de processos com limiar de similaridade de 80%, em conjunto com os testes de Kruskal Wallis e teste post-hoc de Dunn-Bonferroni respectivamente. A partir destes resultados, observa-se que: (1) Há diferença estatística da qualidade média de Recall, Precision e F1 Score dos modelos (cf. Tabela 50) dentre as medidas de similaridade utilizadas no agrupamento de instâncias de processos; e (2) M5 resultou em Recall, Precision e F1 Score estatisticamente superiores as demais medidas de similaridade.

Tabela 49 - Qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos com limiar de similaridade de 80% para *log* de eventos real de multas de tráfego rodoviário

Medida de Similaridade		n	Recall					Precision					FScore				
			mean	sd	min	med	max	mean	sd	min	med	max	mean	sd	min	med	max
M1	Similaridade de rótulos [39]	8	0.637	0.096	0.543	0.596	0.800	0.778	0.179	0.537	0.821	1.000	0.688	0.089	0.557	0.696	0.807
M2	Comparação da dependência de grafos [9]	30	0.771	0.110	0.550	0.754	1.000	0.927	0.095	0.633	0.953	1.000	0.838	0.090	0.649	0.837	1.000
M3	Estimativa de características [45]	5	0.527	0.163	0.364	0.499	0.800	0.561	0.269	0.320	0.400	0.931	0.527	0.182	0.378	0.428	0.779
M4	Distância da edição de grafos e similaridade de rótulos [48]	23	0.752	0.110	0.466	0.765	1.000	0.862	0.149	0.529	0.951	1.000	0.797	0.109	0.573	0.800	1.000
M5	Relação de transições adjacentes [43]	137	**0.953**	0.072	0.615	1.000	1.000	**0.940**	0.142	0.229	1.000	1.000	**0.941**	0.113	0.365	1.000	1.000
Kruskal-Wallis			p-value = 2.2e-16					p-value = **1.346e-07**					p-value = 2.522e-16				

Tabela 50 - Teste *post-hoc* de *Dunn-Bonferroni* com comparação par a par das medidas de similaridade aplicado nos resultados de qualidade dos modelos de processos obtidos no agrupamento de instâncias de processos: limiar de similaridade de 80%, *log* de eventos reais de multas de tráfego rodoviário.

Comparação	p-value		
	Recall	Precision	Fscore
M1 - M2	0.2949637	0.076526594	**0.09449331**
M1 - M3	0.8321057	0.450272838	0.7352108
M2 - M3	0.2481807	**0.0202014**	**0.09792235**
M1 - M4	0.3302471	0.235722756	0.2568229
M2 - M4	0.7445277	0.409841655	0.4326046
M3 - M4	0.3435879	**0.0664188**	0.2124915
M1 - M5	**3.81827E-07**	**0.0016922**	**2.86801E-06**
M2 - M5	**1.97608E-10**	**0.0197353**	**1.54295E-06**
M3 - M5	**1.69018E-05**	**0.0009413**	**3.32889E-05**
M4 - M5	**3.41363E-10**	**0.0023291**	**1.72927E-07**

5.8.2 Síntese dos resultados

O processo de agrupamento de instâncias de processos foi aplicado em um *log* de eventos reais de um processo de gerenciamento multas de tráfego rodoviário. Nos limiares de similaridades de 60% e 80%, M5 mostrou superioridade estatística nas três métricas de qualidade. Desta forma, verifica-se que M5 obteve o melhor desempenho no agrupamento de instâncias de processos.

5.9 Análise dos Resultados

A Tabela 51 apresenta um resumo das análises realizadas nas subseções anteriores. Para cada *log* de eventos e limiar de similaridade são apresentadas as medidas de similaridade, que em conjunto com o agrupamento de instâncias de processos geraram modelos de processos com maior desempenho. As linhas destacadas em **negrito** com **fundo cinza** representam as medidas que estatisticamente obtiveram os melhores desempenhos em relação as demais.

Tabela 51 - Resumo dos resultados de qualidade de modelos de processo do agrupamento de instâncias de processos aplicado com diferentes medidas de similaridade em *log* de eventos sintéticos e reais

Log de eventos		Limiar	Recall	Precision	F1 Score
Sintético	01	40%	M5	Todas	M5
		60%	M1 - M2 - M3	M1 - M2 - M3	M1 - M2 - M3
		80%	M1 - M3	M1 - M3	M1 - M3
	02	40%	M5	M5	M5
		60%	M1 - M3	M1 - M3	M1 - M3
		80%	M5	M1 - M3	M1 - M3
Real	03	40%	M2 - M5	M2 - M5	M2 - M5
		60%	M2 - M5	M2 - M5	M2 - M5
		80%	M2 - M5	M2 - M5	M2 - M5
	04	40%	M5	M5	M5
		60%	M5	M5	M5
		80%	M5	M5	M5
	05	40%	M5	M5	M5
		60%	M5	M5	M5
		80%	M5	M5	M5
	06	40%	M5	M5	M5
		60%	M5	M5	M5
		80%	M5	M5	M5
	07	40%	M1	M5	M1
		60%	M5	M5	M5
		80%	M1	M2	M1
	08	40%	M5	M5	M5
		60%	M5	M5	M5
		80%	M5	M5	M5

Baseado nos resultados apresentados é possível responder a primeira questão de pesquisa (RQ01), na qual **a utilização de diferentes medidas de similaridade selecionadas não impactou na qualidade do agrupamento de instâncias de processos de modo geral**. Essa conclusão é dada visto que não houve diferença estatística dos resultados em todas as bases de dados utilizadas. Observou-se ainda que o agrupamento de instâncias aplicadas em conjunto com a Mineração de Processos gerou modelos de processos com maior qualidade.

Para os *logs* de eventos que representam processos reais com características semiestruturadas e não estruturadas observa-se predominância dos resultados da medida que analisa as relações de transições adjacentes (M5). Nos *logs* de eventos 03, 04, 05, e 08 observou-se superioridade estatística da medida M5 no limiar de similaridade de 80%.

A segunda questão de pesquisa (RQ02), refere-se se é possível estabelecer medidas de similaridade mais apropriadas para o agrupamento de instancias de processos, em particular, quando aplicadas em conjunto com a Mineração de Processos. Conforme resultados apresentados é possível estabelecer que: **(1)** Em *log* de eventos que representam processos mais estruturados, com a maioria das atividades presentes em todas as instâncias de processos, não é possível verificar estatisticamente uma medida de similaridade mais apropriada; **(2)** Em *log* de eventos de processos reais com características não estruturadas, apesar de não haver diferença estatística dos resultados em todos os *logs* de eventos utilizados, observa-se que M5 foi a medida de similaridade que mais obteve resultados estatisticamente superiores dentre as medidas de similaridade utilizadas.

5.10 Limitações da Pesquisa

O estudo realizado apresentou limitações referente a técnica de descoberta empregada em conjunto com o agrupamento de instâncias de processos. Neste caso, foi utilizado o *Inductive Miner*, pois ele é considerado o estado da arte no momento. Para validar a generalização dos resultados, outras técnicas poderiam ser empregadas, como o *Heuristic Miner* [19]*, Genetic Miner* [23] e etc.

Outra limitação apresentada, refere-se ao algoritmo de agrupamento de instâncias utilizado. Neste caso, outra técnica como o agrupamento hierárquico também poderia ser utilizada para validar a generalização dos resultados apresentados.

5.11 Considerações sobre o Capítulo

Este capítulo apresentou a análise e discussão detalhada dos resultados com o objetivo de responder as questões de pesquisa elaboradas. Tal análise foi fundamentada nos testes estatísticos de *Kruskal Wallis* e *Dunn-Bonferroni*. Tais testes foram aplicados sobre os resultados obtidos no experimento. Ao final foi apresentado um resumo e limitações da pesquisa.

CAPÍTULO 6 - CONCLUSÕES

A Mineração de Processos é uma abordagem relativamente recente que permite a observação da execução de processos com base em dados de eventos reais. E para tal, estuda-se métodos computacionais para viabilizar diagnósticos realistas em tempos aceitáveis com alta qualidade. Tais métodos, em particular, são adaptados e/ou desenvolvidos para: (1) descobrir e gerar modelos de processos reais sem conhecimento *a priori*; (2) verificar a conformidade de um modelo de processo teórico por meio do modelo descoberto; (3) aplicar e estender um modelo de processo por meio de informações de desempenho, gargalos, recursos e custos [6] com o objetivo de trazer novas perspectivas de análises.

As técnicas de descoberta de modelos de processos possuem problemas para lidar com *logs* de eventos provenientes de processos flexíveis. Isto resulta geralmente em modelos de processos difíceis de compreender devido ao elevado número de atividades e transições, conhecidos também como modelos espaguete. Uma forma de melhorar os resultados da Mineração de Processo é aplicar técnicas para agrupar as instâncias de processos presente no *log* de eventos. Isto deve ser feito como etapa anterior a etapa de descoberta de modelos de processos para gerar modelos de processos mais estruturados quando comparado com o todo.

A escolha da medida de similaridade empregada no processo de agrupamento impacta diretamente na qualidade dos modelos gerados. Isto foi mostrado por meio do estudo ora realizado para verificar se: (1) as medidas de similaridade impactam na qualidade do agrupamento de instâncias de processos complexos—i.e., semiestruturados ou não-estruturados; (2) é possível estabelecer medidas de similaridade mais apropriadas para o agrupamento de instâncias de processos. Para condução de tal estudo foi realizado um experimento, no qual aplicou-se diferentes medidas de similaridade usando um algoritmo incremental de agrupamento de instâncias de processo com a utilização da técnica de descoberta de modelos de processo *Inductive Miner*.

A partir de análise estatística dos dados coletados **não** foi possível concluir que **as medidas de similaridade** selecionadas **impactam na qualidade dos modelos de processos gerados no agrupamento de instâncias de processo**. Desta forma, não foi possível definir quais medidas de similaridades são as mais apropriadas em âmbito geral, visto que não houve predominância dos resultados em todas os *logs* de eventos. Apesar disto, notou-se que: em *log* de eventos de processos reais com características não estruturadas, apesar de não haver diferença estatística dos resultados em todos os *logs* de eventos utilizados, observou-se superioridade da medida que analisa as relações de transições adjacentes. Por fim, observou-se que a aplicação da tarefa de agrupamento de instâncias de processos em processos não estruturados como etapa antecessora da Mineração de Processos permite obter modelos de processos com qualidade superior.

Como trabalhos futuros, pode-se (1) ampliar o objeto deste estudo com a utilização de diferentes técnicas de descoberta de modelos de processos e diferentes algoritmos de agrupamentos de dados. Desta forma seria possível verificar a generalização dos resultados obtidos neste estudo. Por fim, (2) como visto no referencial teórico, as medidas de similaridade possuem diversas aplicações, como por exemplo: a busca de modelos de processos em repositórios, agrupamentos e verificação de conformidade de modelos de processos. Desta forma, pode-se avaliar a utilização de medidas de similaridade na conformidade de processos em conjunto com a mineração de processos, e em seguida, comparar a eficácia com outras abordagens de conformidade.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] M. Song, C. W. Günther, and W. M. P. Van Der Aalst, “Trace clustering in process mining,” *Lect. Notes Bus. Inf. Process.*, vol. 17 LNBIP, pp. 109–120, 2009.

[2] M. Song, H. Yang, S. H. Siadat, and M. Pechenizkiy, “A comparative study of dimensionality reduction techniques to enhance trace clustering performances,” *Expert Syst. Appl.*, vol. 40, no. 9, pp. 3722–3737, 2013.

[3] M. Becker and R. Laue, “A comparative survey of business process similarity measures,” *Comput. Ind.*, vol. 63, no. 2, pp. 148–167, 2012.

[4] T. Thaler, S. Ternis, P. Fettke, and P. Loos, “A Comparative Analysis of Process Instance Cluster Techniques,” *Proc. der 12. Int. Tagung Wirtschaftsinformatik (WI 2015)*, pp. 423–437, 2015.

[5] B. F. A. Hompes, J. Buijs, W. M. P. van der Aalst, P. M. Dixit, and J. Buurman, “Discovering Deviating Cases and Process Variants Using Trace Clustering,” *Proc. 27th Benelux Conf. Artif. Intell. (BNAIC), Novemb.*, pp. 5–6, 2015.

[6] W. van der Aalst, *Process Mining*, vol. 5. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016.

[7] W. M. P. van der Aalst, “Extracting Event Data from Databases to Unleash Process Mining,” *BPM - Driv. Innov. a Digit. World SE - 8*, pp. 105–128, 2015.

[8] S. Santini and R. Jain, “Similarity measures,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 9, pp. 871–883, 1999.

[9] J.-Y. J. Y. Jung, J. Bae, and L. Liu, “Hierarchical business process clustering,” *Proc. - 2008 IEEE Int. Conf. Serv. Comput. SCC 2008*, vol. 2, no. 12, pp. 613–616, 2008.

[10] R. Lu and S. Sadiq, “On the Discovery of Preferred Work Practice Through Business Process Variants,” in *Conceptual Modeling - ER 2007: 26th International Conference on Conceptual Modeling, Auckland, New Zealand, November 5-9, 2007. Proceedings*, C. Parent, K.-D. Schewe, V. C. Storey, and B. Thalheim, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 165–180.

[11] R. Lu and S. Sadiq, “Business Process Management,” *Bus. Process Manag. 2006*, vol. 4102, no. 1, pp. 426–431, 2006.

[12] W. Sadiq and M. E. Orlowska, “Analyzing process models using graph reduction techniques,” *Inf. Syst.*, vol. 25, no. 2, pp. 117–134, 2000.

[13] J. De Weerdt, M. De Backer, J. Vanthienen, and B. Baesens, “A multi-dimensional quality assessment of state-of-the-art process discovery algorithms using real-life event logs,” *Inf. Syst.*, vol. 37, no. 7, pp. 654–676, 2012.

[14] A. Rozinat and W. M. P. van der Aalst, “Conformance checking of processes based on monitoring real behavior,” *Inf. Syst.*, vol. 33, no. 1, pp. 64–95, 2008.

[15] W. van der Aalst, “Process mining: discovering and improving Spaghetti and Lasagna processes,” *2011 IEEE Symp. Comput. Intell. Data Min.*, no. c, pp. 1–7, 2011.

[16] D. Lin, “An Information-Theoretic Definition of Similarity,” *Proc. ICML*, pp. 296–304, 1998.

[17] W. M. P. Van Der Aalst, A. J. M. M. Weijters, and L. Maruster, “Workflow Mining: Which Processes can be Rediscovered,” *BETA Work. Pap. Ser. WP 74*, p. 25, 2002.

[18] T. Basten and W. M. P. Van Der Aalst, “Inheritance of behavior,” *J. Log. Algebr. Program.*, vol. 47, no. 2, pp. 47–145, 2001.

[19] a. J. M. M. Weijters and W. M. P. van der Aalst, “Rediscovering Workflow Models from Event-Based Data using Little Thumb,” *Integr. Comput. Eng.*, vol. 10, pp. 151–162, 2003.

[20] R. Dijkman, M. Dumas, B. Van Dongen, R. Krik, and J. Mendling, “Similarity of business process models: Metrics and evaluation,” *Inf. Syst.*, vol. 36, no. 2, pp. 498–516, 2011.

[21] A. J. M. M. Weijters, W. M. P. Van Der Aalst, and A. K. A. De Medeiros, “Process Mining with the Heuristics Miner Algorithm,” *Tech. Univ. Eindhoven, Tech. Rep. WP*, vol. 166, pp. 1–34, 2006.

[22] C. W. Günther and W. M. P. van der Aalst, “Fuzzy Mining – Adaptive Process Simplification Based on Multi-perspective Metrics,” *Bus. Process Manag. - Lect. Notes Comput. Sci.*, vol. 4714, pp. 328–343, 2007.

[23] W. M. P. van der Aalst, A. K. A. de Medeiros, and A. J. M. M. Weijters, “Genetic Process Mining,” *Appl. Theory Petri Nets 2005, …*, no. i, pp. 48–69, 2005.

[24] J. M. van der Werf, B. F. van Dongen, C. A. J. Hurkens, K. M. van Hee, and A. Serebrenik, “Process Discovery using Integer Linear Programming,” *Appl. Theory Petri Nets*, vol. 5062, no. 3–4, pp. 368–387, 2008.

[25] E. Lamma, P. Mello, M. Montali, F. Riguzzi, and S. Storari, “Inducing Declarative Logic-Based Models from Labeled Traces,” *Bus. Process Manag.*, vol. 4714 LNCS, pp. 344–359, 2007.

[26] M. V. Zelkowitz and D. R. Wallace, “Experimental models for validating technology,” *Computer (Long. Beach. Calif).*, vol. 31, no. 5, pp. 23–31, 1998.

[27] M. La Rosa, W. M. P. Van Der Aalst, M. Dumas, and F. P. Milani, “Business Process Variability Modeling: A Survey,” *ACM Comput. Surv.*, vol. 50, no. 1, p. 2:1--2:45, 2017.

[28] A. Lindsay, D. Downs, and K. Lunn, “Business process - attempts to find a definition,” *Inf. Softw. Technol.*, vol. 45, pp. 1015–1019, 2003.

[29] R. Bose and W. Van Der Aalst, “Context Aware Trace Clustering: Towards Improving Process Mining Results.,” *Sdm*, pp. 401–412, 2009.

[30] J. De Weerdt, M. De Backer, J. Vanthienen, and B. Baesens, “A robust F-measure for evaluating discovered process models,” *IEEE SSCI 2011 Symp. Ser. Comput. Intell. - CIDM 2011 2011 IEEE Symp. Comput. Intell. Data Min.*, pp. 148–155, 2011.

[31] K. Muralidharan, *Six sigma for organizational excellence: A statistical approach*, vol. 15, no. 5. 2015.

[32] W. M. P. van der Aalst, A. H. M. ter Hofstede, and M. Weske, “Business Process Management: A Survey,” in *Business Process Management: International Conference, BPM 2003 Eindhoven, The Netherlands, June 26--27, 2003 Proceedings*, W. M. P. van der Aalst and M. Weske, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 1–12.

[33] K. Vergidis, A. Tiwari, and B. Majeed, “Business Process Analysis and Optimization: Beyond Reengineering,” *IEEE Trans. Syst. Man, Cybern. Part C (Applications Rev.*, vol. 38, no. 1, pp. 69–82, 2008.

[34] S. J. J. Leemans, D. Fahland, and W. M. P. van der Aalst, “Scalable process discovery and conformance checking,” *Software and Systems Modeling*, pp. 1–33, 2016.

[35] J. De Weerdt, S. Vanden Broucke, J. Vanthienen, and B. Baesens, “Active trace clustering for improved process discovery,” *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 12, pp. 2708–2720, 2013.

[37] A. K. Jain, M. N. Murty, P. J. Flynn, C. Methodologies, and I. Storage, “Data Clustering : A Review,” vol. 1, no. 212, 2000.

[38] A. Adriansyah, B. F. Van Dongen, and W. M. P. Van Der Aalst, “Conformance checking using cost-based fitness analysis,” in *Proceedings - IEEE International Enterprise Distributed Object Computing Workshop, EDOC*, 2011, pp. 55–64.

[39] R. Akkiraju and A. Ivan, “Discovering business process similarities: An empirical study with SAP best practice business processes,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6470 LNCS, pp. 515–526, 2010.

[40] J. Bae, L. Liu, J. Caverlee, L.-J. Zhang, and H. Bae, “Development of Distance Measures for Process Mining, Discovery and Integration,” *Int. J. Web Serv. Res.*, vol. 4, no. 4, pp. 1–17, 2007.

[41] D. Grigori, J. C. Corrales, M. Bouzeghoub, and A. Gater, “Ranking BPEL processes for service discovery,” *IEEE Trans. Serv. Comput.*, vol. 3, no. 3, pp. 178–192, 2010.

[42] Y. Lu, H. Yu, Z. Ming, and H. Wang, “A similarity measurement based on structure of Business Process,” in *2016 IEEE 20th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 2016, pp. 498–503.

[43] H. Zha, J. Wang, L. Wen, C. Wang, and J. Sun, “A workflow net similarity measure based on transition adjacency relations,” *Comput. Ind.*, vol. 61, no. 5, pp. 463–471, 2010.

[44] M. Ehrig, A. Koschmider, and A. Oberweis, “Measuring Similarity Between Semantic Business Process Models,” in *Proceedings of the Fourth Asia-Pacific Conference on Comceptual Modelling - Volume 67*, 2007, pp. 71–80.

[45] Z. Yan, R. Dijkman, and P. Grefen, *Fast business process similarity search with feature-based similarity estimation*, vol. 6426 LNCS, no. PART 1. 2010.

[46] M. Minor, A. Tartakovski, and R. Bergmann, “Representation and Structure-Based Similarity Assessment for Agile Workflows,” *Case-Based Reason. Res. Dev.*, pp. 224–238, 2007.

[47] K. Huang, Z. Zhou, Y. Han, G. Li, and J. Wang, “An algorithm for calculating process similarity to cluster open-source process designs,” *Grid Coop. Comput. 2004Workshops*, no. 60173018, pp. 107–114, 2004.

[48] M. La Rosa, M. Dumas, R. Uba, and R. Dijkman, “Merging Business Process Models,” *Move to Meaningful Internet Syst. (OTM 2010)*, pp. 96–113, 2010.

[49] M. Kunze and M. Weske, “Metric trees for efficient similarity search in large process model repositories,” in *Lecture Notes in Business Information Processing*, 2011, vol. 66 LNBIP, pp. 535–546.

[50] C. Li, M. Reichert, and A. Wombacher, “On measuring process model similarity based on high-level change operations,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2008, vol. 5231 LNCS, pp. 248–264.

[51] J. Bae, J. Caverlee, L. Liu, and H. Yan, “Process mining by measuring process block similarity,” *Bus. Process Manag. Work. Lect. Notes Comput. Sci.*, vol. 4103, pp. 141–152, 2006.

[52] J. Bae, L. Liu, B. J.a, L. L.b, C. J.b, and R. W.B.b, “Process mining, discovery, and integration using distance measures,” *Proc. - ICWS 2006 2006 IEEE Int. Conf. Web Serv.*, pp. 479–486, 2006.

[53] M. Weidlich, J. Mendling, and M. Weske, “Efficient consistency measurement based on behavioral profiles of process models,” *IEEE Trans. Softw. Eng.*, vol. 37, no. 3, pp. 410–429, 2011.

[54] A. Wombacher and M. Rozie, “Evaluation of workflow similarity measures in service discovery,” *Serv. Oriented Electron. Commer.*, vol. 7, no. 26, pp. 51–71, 2006.

[55] K. Gerke, J. Cardoso, and A. Claus, “Measuring the compliance of processes with reference models,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2009, vol. 5870 LNCS, no. PART 1, pp. 76–93.

[56] J. Wang, T. He, L. Wen, N. Wu, A. H. M. Ter Hofstede, and J. Su, “A behavioral similarity measure between labeled Petri nets based on principal transition sequences (short paper),” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010, vol. 6426 LNCS, no. PART 1, pp. 394–401.

[57] A. K. Alves de Medeiros, W. M. P. van der Aalst, and A. J. M. M. Weijters, “Quantifying process equivalence based on observed behavior,” *Data Knowl. Eng.*, vol. 64, no. 1, pp. 55–74, 2008.

[58] S. J. J. Leemans, D. Fahland, and W. M. P. Van Der Aalst, “Discovering block-structured process models from event logs - A constructive approach,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7927 LNCS, pp. 311–329, 2013.

[59] P.-N. Tan, M. Steinbach, and V. Kumar, “Chap 8 : Cluster Analysis: Basic Concepts and Algorithms,” *Introd. to Data Min.*, p. Chapter 8, 2006.

[60] F. Mannhardt, M. De Leoni, H. A. Reijers, and W. M. P. Van Der Aalst, “Data-driven process discovery - Revealing conditional infrequent behavior from event logs,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2017, vol. 10253 LNCS, pp. 545–560.

[61] B. F. Van Dongen, “Real-life event logs - Hospital log.” Eindhoven University of Technology, p. , 2011.

[62] J. C. A. M. Buijs, “Receipt phase of an environmental permit application process (‘WABO’), CoSeLoG project.” Eindhoven University of Technology, p. , 2014.

[63] F. Mannhardt, “Sepsis Cases - Event Log.” Eindhoven University of Technology, p. , 2016.

[64] B. F. Van Dongen, “BPI Challenge 2012.” Eindhoven University of Technology, p. , 2012.

[65] M. De Leoni and F. Mannhardt, “Road Traffic Fine Management Process.” Eindhoven University of Technology, p. , 2015.