MANA: IDENTIFICAÇÃO, MINERAÇÃO, ANÁLISE E REENGENHARIA DE PROCESSOS DE NEGÓCIO

Pedro Miguel Esposito

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia de Sistemas e Computação.

Orientador: Jano Moreira de Souza

Rio de Janeiro

Agosto de 2012

MANA: IDENTIFICAÇÃO, MINERAÇÃO, ANÁLISE E REENGENHARIA DE PROCESSOS DE NEGÓCIO

Pedro Miguel Esposito

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

_____

Prof. Jano Moreira de Souza, D.Sc.

_____

Prof. Guilherme Horta Travassos, D.Sc.

_____

Prof. Cirano Iochpe, D.Sc.

RIO DE JANEIRO, RJ – BRASIL

AGOSTO DE 2012

Esposito, Pedro Miguel

Mana: Identificação, Mineração, Análise e Reengenharia de Processos de Negócio/ Pedro Miguel Esposito. – Rio de Janeiro: UFRJ/COPPE, 2012.

XIV, 134 p.: il.; 29,7 cm.

Orientador: Jano Moreira de Souza

Dissertação (mestrado) – UFRJ/ COPPE/ Programa de Engenharia de Sistemas e Computação, 2012.

Referencias Bibliográficas: p. 127-134.

1. Gerenciamento de Processos de Negócio. 2. Mineração de Processos. 3. Processos Desestruturados. I. Souza, Jano Moreira de. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III Título.

À minha família.

AGRADECIMENTOS

Agradeço primeiramente à minha mãe, Regina, pelo amor, dedicação e suporte em todas as esferas de minha vida; ao meu pai, Wilson, pelo carinho, confiança em meus estudos e por fornecer a base para meu crescimento pessoal e profissional; e à minha irmã, Ana, pelo carinho e auxílio em tantas situações vividas.

Agradeço aos professores que me acompanharam durante toda a jornada na UFRJ. Um agradecimento especial ao professor Jano de Souza, pela orientação, apoio e dedicação ao longo dos últimos três anos, sem o qual este trabalho não teria sido possível. Agradeço ainda aos professores Guilherme Travassos e Cirano Iochpe, por terem aceitado participar da minha banca de defesa de mestrado.

Agradeço ao Marco Vaz, que considero como um segundo pai, pela atenção e auxílio em decidir os caminhos certos a tomar nos níveis acadêmico, pessoal e profissional; sem ele, este trabalho também não teria sido possível. Agradeço ao Sérgio Rodrigues, pelo auxílio durante o mestrado. Agradeço à Patrícia Leal e à Ana Paula Rabello, pelo suporte em várias situações durante esse período.

Agradeço a todos os amigos pela ajuda e carinho, e por estarem sempre lá quando preciso. Em especial, aos amigos Alexandre, Carlos, Emerson, Gustavo, Jonas, Júlia, Rafael, Renan, Thaís e Wagner, que me acompanharam e influenciaram diretamente durante esta jornada, e que tornaram o fundão um ambiente difícil de ser superado. Aos amigos Barreto, Bomfim, Bravo, Conrado, Fernanda, Luana, Lyana, Mariana, Nathalia, Raphael, Rodrigo e Vinícius, por todas as situações vividas e pela compreensão em momentos em que estive ausente durante o mestrado. Agradeço aos novos e antigos amigos da Petrobras, junto dos quais se inicia uma nova jornada.

Agradeço à Universidade Federal do Rio de Janeiro e ao Ministério do Planejamento, Orçamento e Gestão, por fornecerem os dados que viabilizaram as provas de conceito realizadas neste trabalho. Agradeço ainda à COPPE e ao CNPQ, pelo auxílio financeiro sem o qual não teria sido possível me dedicar ao mestrado.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

MANA: IDENTIFICAÇÃO, MINERAÇÃO, ANÁLISE E REENGENHARIA DE PROCESSOS DE NEGÓCIO

Pedro Miguel Esposito

Agosto/2012

Orientador: Jano Moreira de Souza

Programa: Engenharia de Sistemas e Computação

A área de mineração de processos tem o objetivo de superar limitações da modelagem de processos tradicional, utilizando trilhas de auditoria extraídas de sistemas de informação. Diversas técnicas de mineração têm sido propostas na literatura técnica para lidar com processos desestruturados, sendo em sua maioria baseadas em algoritmos de clusterização. As abordagens existentes que suportam tais técnicas, porém, assumem que seja possível extrair previamente um conjunto de instâncias relacionadas. Isso não é a realidade em muitos sistemas que suportam processos desestruturados, que podem possuir tipos de processo genéricos ou permitir a entrada de dados em campos de texto livre. O método MANA foi desenvolvido para lidar com este problema, utilizando apoio ferramental com uma base de dados padrão como ponto de partida para a seleção de instâncias. Dessa forma, o analista é capaz de explorar os dados existentes, identificar instâncias relacionadas e aprimorar seu conhecimento a respeito do funcionamento da organização. Cada conjunto de instâncias de processo selecionado para a mineração pode ser iterativamente refinado até que o modelo de processo desejado seja obtido. A abordagem inclui ainda a análise de desempenho de um processo, através da animação de modelos e de relatórios de desempenho, atividades de reengenharia do processo e a reavaliação do sistema de informação de origem.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

MANA: IDENTIFICATION, MINING, ANALYSIS AND REENGINEERING OF BUSINESS PROCESSES

Pedro Miguel Esposito

August/2012

Advisor: Jano Moreira de Souza

Department: Computer Science Engineering

The process mining field has the goal of overcoming limitations from traditional process modeling techniques, through the use of audit trails extracted from information systems. Several mining techniques have been proposed in the technical literature to deal with unstructured processes, being mostly based on clustering algorithms. Current approaches that support these techniques, however, assume the previous existence of a group of related instances. This is not the reality for several information systems that support unstructured processes, since they may have generic process types or allow the input of data through free text fields. The MANA method was developed to deal with this issue, through tool support and a standard database as the initial phase for instance selection. This approach allows the analyst to explore existing data, identify related instances and gain knowledge about how the organization works. Each instance group can then be iteratively mined and refined until the desired process model can be achieved. The method also includes the performance analysis of a process, through model animations and performance reports, the reengineering of the process and the reevaluation of the source information system.

ÍNDICE

LISTAGEM DE FIGURAS

Capítulo 1 – Introdução

Neste capítulo, o trabalho realizado é contextualizado em relação à área de gerenciamento de processos de negócio. A motivação introduz a importância da modelagem de processos e da descoberta de modelos através de técnicas de mineração de processos. O problema estudado é apresentando, relacionando-se à dificuldade de analisar processos de negócio desestruturados utilizando as abordagens de mineração de processos existentes atualmente na literatura técnica e na indústria. O capítulo inclui ainda uma visão geral da abordagem desenvolvida, contendo a hipótese e os objetivos principais deste trabalho. Finalmente, são descritos o método de pesquisa utilizado e a organização geral do texto da dissertação.

1.1 Motivação

Toda empresa deseja ser flexível, competitiva, inovadora, eficiente e lucrativa. Porém, a grande maioria delas funciona da forma inversa, sofrendo com problemas de rigidez, ineficiência e baixa satisfação de seus consumidores. A grande causa desta aparente contradição não está relacionada a problemas gerenciais ou à falta de motivação de seus empregados, mas na maneira como cada empresa se organiza e executa suas atividades. Para se tornar competitiva, então, uma organização precisa avaliar seu funcionamento interno, identificando os problemas em seus processos que a impedem de atingir seus objetivos estratégicos. Para atingir a eficiência, uma empresa tradicional deve ser capaz de se reinventar, organizando-se em torno de seus processos e focando no resultado gerado para seu consumidor (Hammer e Champy 1994).

No Brasil, com a obrigatoriedade de elaboração do Plano Diretor de Tecnologia da Informação, estabelecido a partir da Instrução Normativa 04 (Brasil 2008), as organizações públicas iniciaram o processo de descoberta da importância de possuir um planejamento estratégico organizacional, além de um planejamento estratégico para a Tecnologia da Informação. Um dos frutos do trabalho desenvolvido é a percepção da importância da estruturação e modelagem dos processos de negócio. A modelagem de processos é especialmente importante para esse tipo de organização, que são tradicionalmente consideradas ineficientes e contendo grandes deficiências processuais.

Processo pode ser definido como a maneira com que as empresas organizam seus recursos, contendo seu fluxo de trabalho, dividido em atividades, e quem é responsável pela execução de cada uma delas. Os processos de negócio fazem parte do cerne de qualquer instituição, sendo, dessa forma, um importante ativo que precisa ser bem gerenciado e entendido para que ela atinja seus objetivos. Eles precisam ser bem definidos, entendidos e documentados, para que sejam executados sempre de maneira uniforme e consistente. Além disso, seu correto gerenciamento é vital para que o conhecimento sobre a execução das atividades de uma empresa não esteja somente na mente de seus executantes, mas se torne propriedade intelectual da organização (Schedlbauer 2010).

O gerenciamento de processos de negócio, ou BPM (do inglês *Business Process Management*), é a área de conhecimento que engloba todos os conceitos, metodologias e atividades envolvidas no apoio ao ciclo de vida de um processo. Estas atividades são organizadas no chamado ciclo de vida BPM (Weske 2007), que se inicia com a identificação dos processos executados por uma organização, sua explicitação em modelos e a análise destes modelos através de técnicas de validação, simulação e verificação. Os processos podem então ser selecionados para receberem suporte de um sistema de execução de processos, envolvendo atividades de implantação, teste, operação e manutenção. Técnicas de avaliação, como a mineração de processos e o monitoramento de suas atividades possibilitam a melhoria contínua dos processos de negócio, e têm recebido grande atenção nos últimos anos.

O interesse pelo estudo do gerenciamento de processos de negócio tem motivado cientistas de diferentes áreas. Estudiosos de administração de empresas têm como objetivo otimizar o funcionamento interno de organizações, eliminando custos, fomentando a inovação e aumentando o grau de satisfação do consumidor. Na ciência da computação, tanto pesquisadores de métodos formais, que estudam abstrações estruturais de processos, quanto engenheiros de software são atraídos pela área (Weske 2007). Estes têm como objetivo estudar a integração de sistemas de informação pertencentes à organização dentro de um fluxo de processo, além de desenvolver novos sistemas que apoiem a modelagem e a execução robusta dos processos de uma organização.

Uma das primeiras e mais importantes atividades executadas em projetos de BPM é o desenho de processos. Os processos de negócio da empresa são identificados e posteriormente

modelados, geralmente através de notações específicas. Existem diversas linguagens para a modelagem de processos, cada uma suportada por uma ampla gama de ferramentas. A especificação de um processo envolve a identificação das tarefas executadas, seus relacionamentos e as regras de negócio às quais o processo está sujeito. A modelagem de processos é uma atividade intensiva, sendo executada através de técnicas como entrevistas, brainstorming, observação e análise de documentos. Cada processo deve ainda ser cuidadosamente analisado, para que suas deficiências sejam identificadas e tratadas adequadamente. Técnicas de análise de processos incluem simulação, análise estrutural, avaliação de desempenho e discussões com especialistas do domínio. Simulações são especialmente importantes durante a análise de processos, identificando comportamentos indesejados a partir de sua execução automatizada. Cada modelo deve ainda ser validado por suas partes interessadas através da realização de *workshops* (van der Aalst et al. 2003b).

A modelagem de processos tradicional é, porém, uma atividade altamente custosa, exigindo uma grande mobilização de recursos como pessoal, tempo e dinheiro. Muitas organizações, principalmente aquelas sem fins lucrativos, encontram dificuldades em motivar uma empreitada desse porte, tornando projetos de BPM impraticáveis (Greco et al. 2006). Abordagens tradicionais são ainda fortemente baseadas em entrevistas com as partes interessadas do processo. Cada pessoa, porém, pode ter uma visão tendenciosa que não corresponde exatamente à realidade. Isso pode resultar em erros de modelagem, idealizações e generalizações excessivas (van der Aalst 2011). As ineficiências do processo, um dos principais aspectos a identificar, podem passar amplamente despercebidas pelo analista.

A mineração de processos, área de pesquisa que tem recebido grande atenção nos últimos anos, surge na tentativa de solucionar as deficiências existentes na modelagem tradicional. Seu conceito principal é a utilização de informações de execução de instâncias reais de um processo para analisá-lo. Os dados utilizados são extraídos de logs de eventos registrados por sistemas de informação que apoiam o processo. Assume-se que, para cada atividade executada, foi registrado pelo menos um evento no log. Dessa forma, é possível identificar, para cada instância do processo, suas atividades, o momento em que elas ocorreram e seus executores, dentre outras informações.

As principais técnicas de mineração de processos envolvem a descoberta de modelos de processo, extraindo seu fluxo a partir do log de eventos, como mostra a Figura 1. Ou seja, é feita a engenharia reversa de um modelo de processo a partir de um conjunto de instâncias reais deste processo. A partir de configurações dos diversos algoritmos propostos na literatura, é possível obter diferentes níveis de granularidade no processo. Como a mineração de processos obtém modelos a partir de dados reais de execução, o fluxo modelado é menos sujeito a erros do que aquele obtido através de entrevistas. Sua contestação fica dificultada, permitindo utilizá-lo como prova da necessidade de se iniciar um projeto de reengenharia organizacional. A mineração de processos pode ainda ser utilizada como um ponto de partida para alavancar projetos de BPM em uma empresa, dado que o custo inicial para a obtenção de modelos de processo é muito menor do que com a abordagem tradicional. Ainda que possam necessitar de uma avaliação final pelos participantes do processo, para enriquecer o fluxo identificado e corrigir interpretações equivocadas dos algoritmos, os modelos obtidos através da mineração de processos são um importante insumo que torna mais eficiente e eficaz as demais atividades de um projeto de BPM.

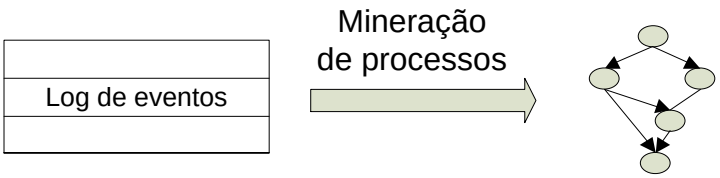

Figura 1 – Descoberta de modelos utilizando técnicas de mineração de processos

A principal ferramenta de mineração de processos existente hoje é o framework ProM (Van Dongen et al. 2005). Ele implementa o estado da arte em técnicas de mineração e análise de processos. Seus mais de 230 plug-ins (da Cruz e Ruiz 2008) permitem analisar diversas perspectivas de um processo e cobrem diversos casos de uso. A entrada de dados no ProM é feita através de arquivos XML, contendo logs de eventos. Ele pressupõe que é possível registrar, para os eventos do log, as atividades executadas, as instâncias relacionadas a cada atividade e o momento em que elas ocorreram. Outras informações podem ser armazenadas, como o responsável pela execução de uma atividade. A construção dos arquivos importados pelo framework pode ser auxiliada por ferramentas externas, como o ProM Import (Günther e van der Aalst 2006) e o Nitro (Fluxicon 2012). A utilização do framework não é trivial, exigindo um bom conhecimento da área de mineração de processos. Para tirar proveito de um

algoritmo, geralmente se faz necessária a leitura do artigo científico descrevendo a técnica. Dessa forma, muitos usuários que poderiam se beneficiar da mineração de processos encontram dificuldade em explorar a ferramenta. O framework ProM será utilizado como comparativo para a abordagem desenvolvida neste trabalho.

1.2 Problema

Embora os algoritmos de descoberta obtenham sucesso quando o processo modelado é razoavelmente bem estruturado, elas falham quando existem muitos casos excepcionais e pouca relação de dependência entre as atividades executadas (van der Aalst e Gunther 2007). Para estes processos desestruturados, a mineração utilizando técnicas tradicionais resulta em modelos em espaguete, altamente complexos e com pouca utilidade para o negócio. Um modelo de processo em espaguete é exemplificado na Figura 2. Este resultado ocorre devido ao grande número de fluxos distintos extraídos do log de eventos. Isso gera um desafio, pois os processos desestruturados são aqueles que mais poderiam se beneficiar de técnicas de modelagem e reengenharia de processos. Nota-se que os modelos em espaguete não estão necessariamente errados; eles somente reproduzem a maneira caótica como o processo é executado.

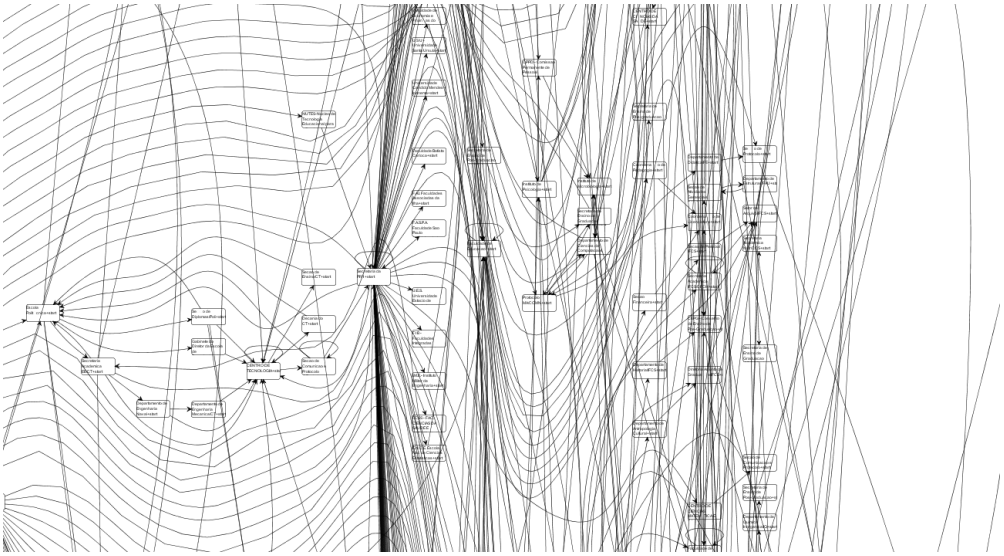

Figura 2 - Exemplo de modelo de processo em espaguete

Para lidar com esse tipo de processo, diversas técnicas têm sido propostas na literatura técnica. A grande maioria segue uma abordagem "dividir para conquistar". Elas procuram

dividir um log de eventos em conjuntos menores de instâncias de processo que possuam grande similaridade entre si, simplificando o problema. A clusterização é feita a partir de algoritmos tradicionais de mineração de dados. Uma abordagem alternativa é utilizar uma abstração de mapas, permitindo controlar o nível de granularidade desejado no modelo. Nesse caso, ao se distanciar do mapa, conjuntos de atividades são clusterizados em um mesmo elemento do modelo.

As abordagens atuais de mineração de processos, porém, são suportadas por ferramentas que exigem a importação de tipos de processo bem definidos. Mesmo o framework ProM, que permite a mineração de processos desestruturados, exige como ponto de partida para seu fluxo de trabalho a importação de um arquivo contendo um log de eventos. Existem ferramentas que auxiliam na criação destes logs, porém elas possuem enfoque no mapeamento entre tipos de dados do sistema de origem para um padrão de log, e não na identificação de instâncias relacionadas.

Muitos sistemas de informação que armazenam dados de processos, porém, não possuem uma separação clara entre tipos de processo. Isso é realidade, por exemplo, em diversos sistemas de protocolo de organizações públicas, que possuem uma classificação fraca de processos e permitem seu detalhamento através de campos de entrada textual livre. A obtenção de resultados satisfatórios a partir das abordagens de clusterização fica dificultada caso os dados de origem sejam muito heterogêneos; a efetividade dos algoritmos, a quantidade de clusters desejada, e o que cada cluster representa na prática são alguns dos problemas encontrados. Além disso, dados importantes a respeito de cada instância de processo, que podem ser utilizados para se obter conhecimento sobre o funcionamento da organização, ficam perdidos na conversão entre sistemas.

Para que um projeto de mineração de processos tenha sucesso, a identificação de instâncias relacionadas deve ser realizada de maneira exploratória (van der Aalst e Gunther 2007). O usuário deve ser capaz de analisar iterativamente a base de dados, aprendendo com seus erros para identificar um processo e obter um modelo de processo considerado razoável. Dessa forma, a identificação de instâncias de processo relacionadas deveria ser uma atividade importante em um método de mineração e análise de processos, e integrada à ferramenta utilizada pelo analista. A visão completa dos processos executados pela organização auxilia

na compreensão de como o negócio é executado e na priorização dos processos que necessitam de reengenharia.

Dessa forma, esta pesquisa foi desenvolvida com a hipótese de que *em uma organização que possua dados de processos armazenados em um sistema de informação, sem classificação detalhada e sem fluxos bem definidos, existem casos em que é possível obter resultados com maior valor para o analista e um maior nível de conhecimento nas atividades de mineração destes processos, em relação às abordagens existentes, utilizando um método fundamentado na exploração de uma base de dados padrão contendo atributos de instâncias de processo, que permita a seleção incremental de instâncias relacionadas.*

1.3 Método MANA

O objetivo principal desta pesquisa é *definir um método que permita aprimorar as atividades de mineração e análise de processos de negócio desestruturados utilizando uma base de dados padrão contendo atributos de instâncias, integrando atividades de identificação, mineração, análise e reengenharia de processos, com foco na atividade de identificação, permitindo a exploração da base para adquirir conhecimento incremental a respeito dos processos analisados*. Para isso, foi desenvolvido o método MANA. Embora este trabalho tenha sido motivado pela necessidade de modelagem dos processos e reengenharia de sistemas de organizações públicas, a abordagem desenvolvida pode ser utilizada com dados provenientes de qualquer sistema de informação que registre instâncias de processo, o fluxo de atividades de cada instância e atributos que permitam identificar instâncias de um mesmo processo. Outros objetivos deste trabalho, relacionados ao principal, incluem:

- Desenvolver uma ferramenta que suporte o método proposto;
- Auxiliar na identificação de instâncias de processo relacionadas, permitindo a exploração da base padrão, através de buscas e filtros textuais sobre atributos relevantes para o negócio, e utilizando técnicas automatizadas;
- Facilitar o uso de técnicas de mineração de processos por analistas de negócio que não sejam especialistas na área;
- Motivar a reengenharia de processos, facilitando a análise de seu estado atual através de informações visuais.

O método MANA possui enfoque na etapa de identificação de instâncias de processo relacionadas, utilizando atributos semanticamente relevantes para o negócio como assunto, descrição, origem, partes interessadas, ano e unidades participantes. Estes atributos são extraídos do sistema de informação de origem e carregados para a base de dados padrão que suporta a abordagem. Isso permite que o usuário adote uma abordagem exploratória, procurando identificar os processos importantes da organização e as instâncias que se relacionam a ele. Técnicas automatizadas podem auxiliar nesta etapa quando necessário. A exploração de informações relevantes para o processo é importante para que o analista seja capaz de entender em que casos cada fluxo do processo é executado, ampliando seu conhecimento sobre o funcionamento da organização.

A Figura 3 fornece uma visão geral do método MANA. A primeira etapa é a extração de dados a partir de um sistema de informação que registre instâncias de processo para uma base de dados padrão. Na etapa de identificação é construída uma *consulta de processo*. Uma consulta é uma pasta de trabalho, contendo um conjunto de instâncias selecionadas que serão utilizadas para as atividades de modelagem e análise de processos. Uma consulta é construída através de um conjunto de filtros, que são incluídos pelo usuário a partir da exploração das instâncias disponíveis. Por exemplo, o usuário pode desejar incluir todas as instâncias que contenham o termo *contrato* em sua descrição, e que passaram pelo departamento de TI da organização. A busca de informações a partir de pesquisas textuais é um fator importante para que a identificação de processos a partir de dados desestruturados obtenha sucesso.

A etapa de mineração se relaciona à descoberta de modelos de processo. Ela é realizada em cima das instâncias filtradas em uma consulta, utilizando algoritmos de mineração de processos descritos na literatura. Caso as informações a respeito das atividades executadas sejam deficientes na base de dados, é possível modelar o fluxo do processo entre as unidades organizacionais. Isso é importante quando o sistema de origem possui enfoque maior em quem executou uma atividade do que na atividade em si. Nota-se que o modelo de processo resultante da mineração deve ser validado e enriquecido em conjunto com as partes interessadas do processo.

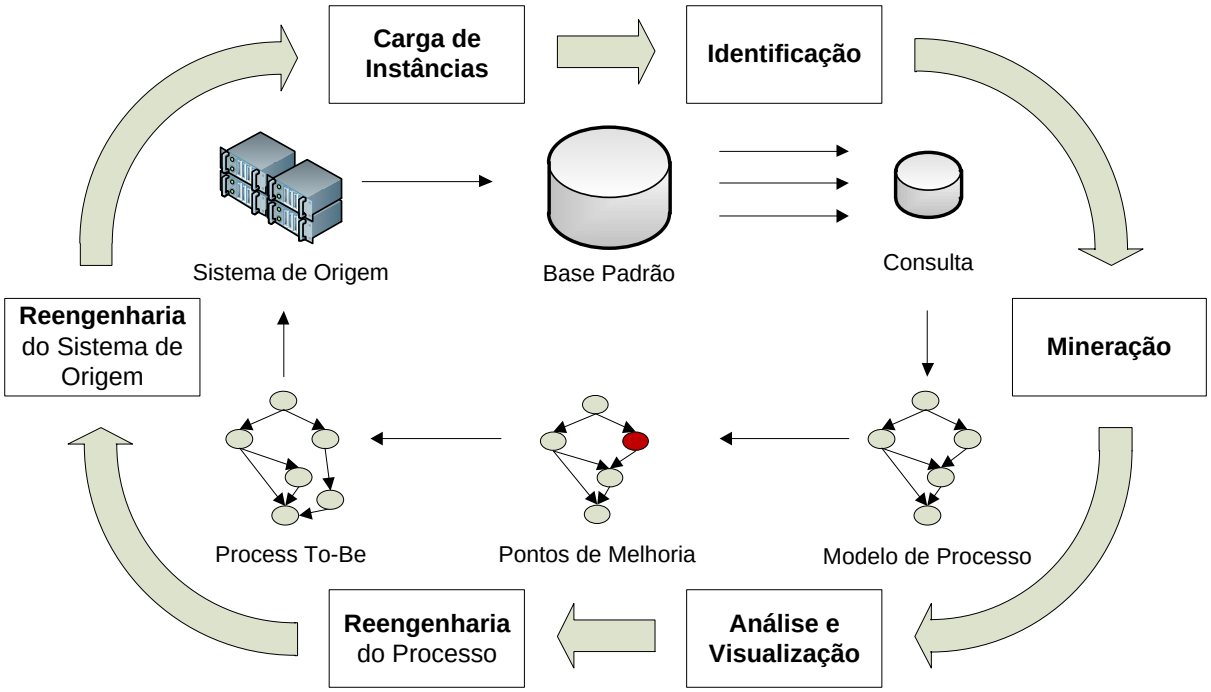

Figura 3 – Visão geral do método MANA

A abordagem proposta dá grande ênfase ao aspecto visual da análise de processos. A animação de processos, por exemplo, permite visualizar o andamento das instâncias de um processo ao longo do tempo, possibilitando uma análise intuitiva do estado atual do processo e a identificação de suas principais deficiências. A ferramenta desenvolvida neste trabalho permite ainda a extração de relatórios e a exibição de gráficos, indicando o desempenho de cada unidade organizacional durante a execução do processo. As atividades de análise permitem que as deficiências do processo sejam identificadas, viabilizando a construção de modelos *to-be* e motivando uma reavaliação do sistema de origem para suportar uma maior estruturação de tipos de processo e do fluxo a ser seguido. Finalmente, novas instâncias podem ser carregadas para a base padrão, permitindo a avaliação contínua dos processos da organização.

Este trabalho inclui duas provas de conceito que possuem o objetivo de apresentar situações reais onde o método MANA é capaz de aprimorar a análise de processos de negócio desestruturados, permitindo ao analista obter resultados mais precisos e um maior nível de conhecimento a respeito dos processos da organização. A primeira prova de conceito utilizou dados do Sistema de Controle de Processos e Documentos do Ministério do Planejamento, Orçamento e Gestão, com foco no processo de desfazimento de equipamento de informática

com participação da Secretaria de Logística e Tecnologia da Informação. A segunda prova de conceito utilizou dados do Sistema de Acompanhamento de Processos da Universidade Federal do Rio de Janeiro, com foco no processo de registro dos diplomas de alunos de graduação recém-formados na Universidade.

1.4 Metodologia de Pesquisa

Este trabalho pretende seguir a metodologia de pesquisa abaixo, como descrita por Marconi e Lakatos (2004 apud Rodrigues 2011). Segundo os autores, uma pesquisa atinge seus objetivos através das seguintes etapas:

- Descobrimento do problema;
- Colocação precisa do problema;
- Procura de conhecimentos ou instrumentos relevantes ao problema;
- Tentativa de solução do problema com o auxílio dos meios identificados;
- Produção de novos dados;
- Obtenção da solução;
- Investigação das consequências da solução obtida;
- Prova (comprovação) da solução;
- Correção das hipóteses, teorias, procedimentos ou dados empregados na obtenção da solução incorreta.

O descobrimento do problema se deu pela dificuldade de obter resultados satisfatórios utilizando as abordagens existentes de mineração de processos a partir de dados extraídos de sistemas de informação sem classificação detalhada de instâncias e sem fluxo bem definido. Isso levou à colocação do problema, como apresentado na seção 1.2. Foi feita uma revisão da literatura técnica relacionada aos algoritmos e abordagens existentes atualmente que se propõem a resolver o problema.

Dados foram gerados utilizando as abordagens atuais e a abordagem desenvolvida. O método MANA foi proposto como solução para o problema, tendo sido investigado através da execução de provas de conceito. Provas de conceito foram realizadas para investigação e comprovação de que existem casos em que a abordagem apresentada possui resultados que agregam maior valor à análise de mineração de processos do que as abordagens existentes na literatura técnica. Alguns procedimentos empregados foram corrigidos durante o desenvolvimento do trabalho para se adequarem aos resultados experimentais obtidos.

1.5 Organização do Trabalho

Este capítulo apresentou a motivação deste trabalho, o problema em questão, os objetivos propostos, uma visão geral do método desenvolvido e a metodologia de pesquisa utilizada. Os demais capítulos estão organizados da seguinte maneira:

- O capítulo 2 introduz a área de Gerenciamento de Processos de Negócio (BPM), discutindo a definição de processo, o ciclo de vida BPM, a modelagem de processos e suas notações gráficas, e os conceitos de *Business Process Analysis* (BPA), *Business Intelligence* (BI) e *Business Process Intelligence* (BPI);

- O capítulo 3 apresenta o estado da arte na mineração de processos, discutindo o funcionamento de algoritmos de descoberta e de clusterização de processos;

- O capítulo 4 discute as abordagens existentes relacionadas ao problema apresentado, com foco no framework ProM e no Aris Process Performance Manager;

- O capítulo 5 apresenta o método MANA, detalhando o fluxo de trabalho proposto, além de discutir seus principais conceitos e os diferenciais da abordagem adotada em relação às existentes na literatura. O capítulo introduz ainda a ferramenta desenvolvida para suportar a abordagem proposta;

- O capítulo 6 introduz duas provas de conceito, utilizando dados extraídos de sistemas de acompanhamento de processos do Ministério do Planejamento e da Universidade Federal do Rio de Janeiro. Esses estudos procuram mostrar que existem casos em que

é possível obter resultados melhores com o método proposto do que com as abordagens existentes;

- O capítulo 7 conclui este trabalho, realizando as considerações finais, ressaltando os resultados e contribuições obtidas, expondo as limitações do trabalho desenvolvido e apresentando suas direções futuras.

Capítulo 2 – Gerenciamento de Processos de Negócio

Este capítulo introduz o campo de pesquisa do Gerenciamento de Processos de Negócio, ou BPM (do inglês *Business Process Management*), apresentando definições e metodologias presentes na literatura e que serão utilizadas no decorrer deste trabalho. Os conceitos de processo e de BPM são discutidos de acordo com as definições dadas por alguns dos principais autores da área. O ciclo de vida BPM é introduzido, contextualizando como este trabalho está inserido no decorrer de suas etapas. É dada atenção especial à modelagem de processos e suas principais notações gráficas, com foco no *Business Process Model and Notation* (BPMN), utilizado neste trabalho. O capítulo introduz ainda os principais conceitos da análise de processos e a área relacionada de *Business Intelligence* (BI).

2.1 Definições

Nos anos 70 e 80, as soluções de *software* eram primariamente focadas no armazenamento e recuperação de dados e informações, em abordagens guiadas pela modelagem de dados. Os processos de negócio, dessa forma, precisavam se adaptar à arquitetura de TI existente. No entanto, essa perspectiva tem se alterado para o desenvolvimento de soluções guiadas por processos. Além disso, a tendência se dá na direção do redesenho de processos, ao invés de modelos fixos e bem planejados. Isso permite que os processos de negócio se adaptem à evolução da empresa e do ambiente em que ela está inserida (van der Aalst et al. 2003b).

Antes de definir Gerenciamento de Processos de Negócio, ou BPM (do inglês *Business Process Management*), van der Aalst et al. (2003b) primeiro consideram a definição tradicional de *workflow*, pois a área de BPM surgiu a partir da evolução do gerenciamento de *workflows*. O *Workflow Management Coalition* (WfMC) define *workflow* como: "A automação de processos de negócio, completamente ou em parte, durante a qual documentos, informação e tarefas são repassados de um participante a outro para a execução de ações, de acordo com um conjunto de regras procedurais", e *Workflow Management System* (WFMS) como "Um sistema que define, cria e gerencia a execução de *workflows* através do uso de software, executado em uma ou mais máquina de *workflow*, que são capazes de interpretar a definição do processo, interagir com os participantes do *workflow* e, onde necessário, invocar o uso de ferramentas e aplicativos de TI" (Lawrence 1997).

Segundo van der Aalst et al. (2003b), o conceito de *workflow* é muito restritivo, focando-se na utilização de sistemas de software para apoiar a execução de processos operacionais. Dessa forma, os autores ampliam o conceito, definindo *Business Process Management* como "O suporte a processos de negócio utilizando métodos, técnicas e software para projetar, implementar, controlar e analisar processos operacionais envolvendo humanos, organizações, aplicativos, documentos e outras fontes de informação." Processos estratégicos e que não podem ser explicitados são excluídos dessa definição, amarrada a processos operacionais.

Weske (2007) fornece uma definição similar. Para o autor, processo de negócio é "um conjunto de atividades que são executadas coordenadamente em um ambiente técnico e organizacional. Essas atividades realizam conjuntamente um objetivo do negócio. Cada processo de negócio é executado por uma única organização, mas podendo interagir com processos de outras organizações." Já BPM é definido como os "conceitos, métodos e técnicas que apoiam o projeto, a administração, a configuração a execução e a análise de processos de negócio".

Em Valle e de Oliveira (2009), os autores ressaltam a distinção entre organizações tradicionais com estrutura funcional e organizações cujo foco se dá nas atividades de trabalho executadas por seus processos. A organização vertical dá lugar a uma organização horizontal. Os autores se baseiam na descrição de Davenport (1994 apud Valle e de Oliveira 2009), que define processo como "um conjunto de atividades estruturadas e medidas, destinadas a resultar num produto especificado para um determinado cliente ou mercado. [...] É, portanto, uma ordenação específica de atividades de trabalho no tempo e no espaço, com um começo, um fim e *inputs* e *outputs* claramente identificados: uma estrutura para a ação. [...] Enquanto a estrutura hierárquica é, tipicamente, uma visão fragmentada e estanque das responsabilidades e das relações de subordinação, a estrutura de processo é uma visão dinâmica da forma como a organização produz valor".

Os autores sugerem a representação abaixo para a visão de um processo. Suas entradas são recursos que são transformados ou utilizados para propiciar a transformação. Obedecendo as regras às quais o processo está submetido, estes recursos são processados nas saídas do processo. Um processo está ainda envolvido em um contexto (não presente na figura): a

criação tecnológica, a esfera doméstica e cultural, as estruturas políticas e jurídicas e o mercado.

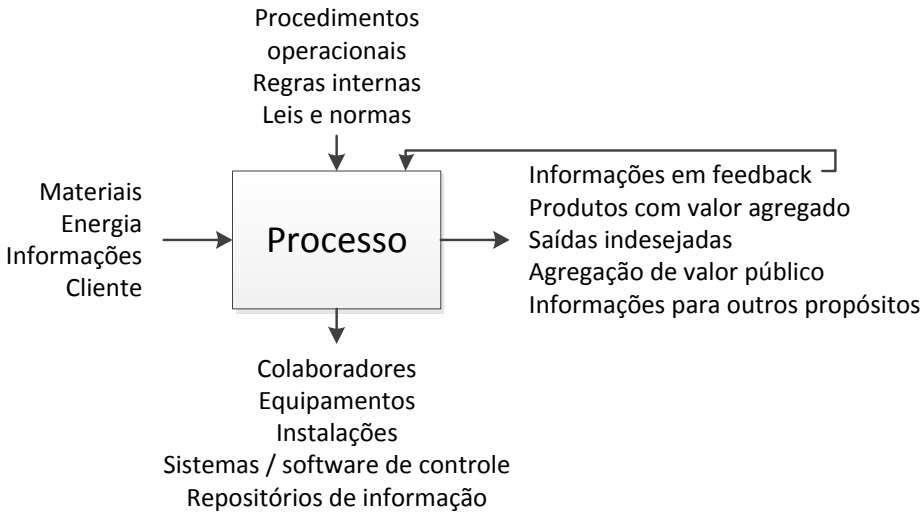

Figura 4 - Modelo para a visão de um processo - adaptado de Valle e de Oliveira (2009)

As atividades que compõem um processo podem ser atividades de sistema, de interação de usuário ou manuais. As atividades manuais, como o envio de uma encomenda, não possuem a participação de sistemas de informação. Atividades de interação de usuário são realizadas por pessoas através da utilização de sistemas de informação, como o preenchimento de um formulário de venda on-line. Atividades de sistema são realizadas inteiramente pelo computador, sem a participação de usuários. Por exemplo, um sistema pode verificar automaticamente informações de uma conta bancária, contanto que ele possua os dados necessários para isso. Algumas vezes, mudanças de estado resultantes de atividades manuais são cadastradas no computador por meio de atividades de interação de usuário (Weske 2007). Isso é muito comum em sistemas de protocolo, como os que serão estudados posteriormente neste trabalho.

Vale ressaltar que, na prática, o termo processo de negócio pode se referir tanto à sua modelagem abstrata quanto a uma instância específica do processo; o mesmo ocorre para suas atividades (Weske 2007). Para evitar confusões este texto utiliza termos específicos, como instância de processo, modelo de processo, tipo de processo e instância de atividade, que serão detalhados na seção 5.2.

Outro conceito importante da área é o de Sistema de Gerenciamento de Processos de Negócio, ou BPMS (do inglês *Business Process Management System*), definido por van der Aalst et al. (2003b) como "um sistema de software genérico que é guiado por desenhos de processo explícitos, para atuar em e gerenciar processos de negócio operacionais. Eles devem suportar, principalmente, representações gráficas de processos estruturados, e permitirem alterações nos processos suportados. Um sistema BPMS permite evitar que processos sejam codificados em soluções personalizadas de software. Dessa forma, o foco se dá na orquestração, permitindo, por exemplo, integrar aplicativos existentes, cuja interação pode ser modificada somente com a adaptação do modelo de processo."

A utilização de um BPMS para a orquestração das atividades dos processos de uma organização contrasta com a abordagem tradicional, onde a coordenação é realizada manualmente (Weske 2007). Como sistemas BPMS são voltados à execução de processos, enquanto este trabalho se relaciona à sua modelagem e análise, eles não serão estudados a fundo. Cabe ressaltar, porém, que os logs de execução utilizados pela mineração de processos podem ser obtidos a partir de sistemas BPMS. Além disso, a etapa de reengenharia, que ocorre após a organização analisar suas deficiências, pode incluir a adoção de um sistema de BPMS para que os processos passem a ser suportados por um sistema ciente de seu fluxo de atividades.

2.2 Ciclo de Vida BPM

Diversos ciclos de vida para as atividades de BPM têm sido propostos na literatura técnica, como os modelos de van der Aalst (2004), Weske (2007), e Valle e de Oliveira (2009). Embora se aproximem, eles diferem na separação das atividades de BPM em etapas. Em Rós et al. (2009), os autores fornecem uma lista abrangente de modelos existentes, e procuram enquadrar as etapas de cada um deles dentro de um mesmo modelo padrão. A definição abaixo se baseia nos ciclos propostos por van der Aalst (2004) e Weske (2007), dois dos principais autores da área de BPM. A sequência das etapas do ciclo de vida não indica necessariamente uma sequência temporal entre as atividades executadas. Vale ressaltar que o foco deste trabalho se dá nas etapas de projeto, análise e avaliação de processos. As etapas de configuração e promoção envolvem a utilização direta de sistemas BPMS, que, como visto anteriormente, são considerados de maneira indireta pela abordagem proposta.

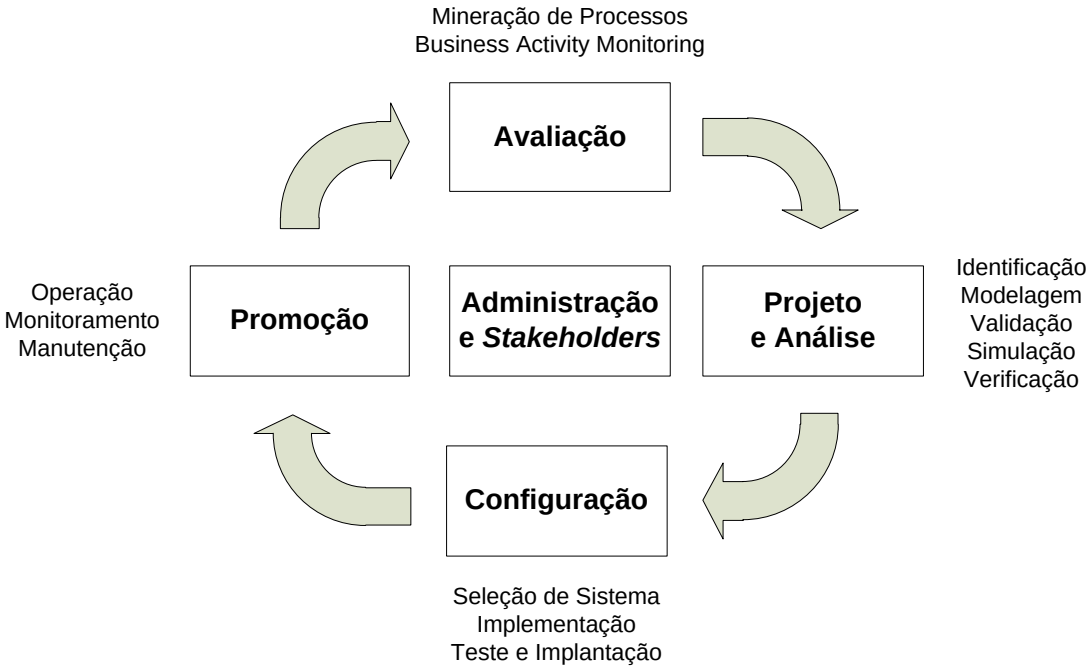

Figura 5 – Ciclo de vida BPM - adaptado de van der Aalst (2004) e Weske (2007)

Na etapa de projeto e análise, os processos de negócio da organização são investigados, analisados e modelados. Com base nas informações coletadas a partir de pesquisas internas, são construídos modelos relacionados às diversas perspectivas do processo, como de fluxo de dados, de fluxo de controle, organizacional, operacional e sócio-técnica. Este conhecimento é utilizado para a configuração de sistemas de informação cientes de processo (van der Aalst 2004). A notação utilizada para os modelos de processo deve permitir a comunicação entre as diversas partes interessadas da empresa (Weske 2007).

Além da modelagem, esta etapa envolve a verificação, a simulação e a validação dos processos modelados. A verificação identifica se um modelo de processo está correto e não possui propriedades indesejadas, como a ocorrência de *deadlocks*. Simulações automatizadas permitem a identificação de comportamentos indesejados resultantes da execução dos modelos de processo. A validação pode ser realizada através de *workshops*, onde as partes interessadas confirmam que o modelo de processo está de acordo com a realidade (Weske 2007). A validação também pode ser executada através de algoritmos de conformidade da mineração de processos, que comparam um modelo com instâncias reais extraídas de um log de execução.

Na fase de configuração, um BPMS é configurado, através da orquestração de outros sistemas que atendam às atividades executadas pelos processos de negócio. Alternativamente, um processo pode ser implementado sem a utilização de sistemas BPMS. Nesse caso, são estabelecidos procedimentos e políticas que devem ser seguidos pelos funcionários da organização (Weske 2007). Os processos devem ser priorizados, através de métricas que avaliem sua necessidade de automação e otimização (Sharon et al. 1997). Os modelos de processo obtidos na fase de projeto devem ser enriquecidos com informações técnicas que suportem sua implantação no ambiente de BPMS escolhido pela organização. Testes devem ser executados, utilizando técnicas provenientes da engenharia de software. Pode ser necessário também realizar treinamentos para os participantes do processo (Weske 2007).

Na fase de promoção (tradução livre do inglês *enactment*), o sistema de informação empresarial implantado é utilizado pela organização. Eventos iniciam instâncias de processo, que passam a ser controladas pelo BPMS, de acordo com os modelos criados na fase de projeto e enriquecidos na fase de configuração. Os processos são monitorados ativamente durante sua execução, permitindo a obtenção do status de cada instância em tempo real (Weske 2007).

A fase de avaliação fecha o ciclo, reiniciando-o, através da avaliação e implantação de melhorias, envolvendo diversos fatores como novas tecnologias, análise de mudanças ambientais, melhoria de desempenho, evolução da empresa, dentre outros. Esta fase inclui a mineração de processos e o *Business Activity Monitoring* (BAM), que serão detalhados nas seções seguintes deste trabalho. Em van der Aalst (2004), o autor ressalta que esta fase não está presente em sistemas de *workflow* tradicionais, que possuem somente suporte básico às fases de projeto e promoção. A fase de avaliação é fortemente relacionada à fase de análise (Weske 2007), sendo importante para a reengenharia de processos. A avaliação de processos possibilita a identificação de problemas e o desenvolvimento de modelos de processo otimizados *to-be*.

2.3 Modelagem de Processos de Negócio

Uma das atividades iniciais executadas durante a implantação do BPM em uma organização é seu projeto, que envolve múltiplos passos. Presente nesta fase está a modelagem de processos, que se refere à utilização de modelos para representar um processo.

A modelagem inclui a identificação e especificação cada processo, com a modelagem de suas atividades, dos relacionamentos entre elas e das regras de negócio relacionadas ao processo. Existem diversas linguagens disponíveis para a modelagem de processos, que se diferenciam em sua semântica, poder expressivo e suporte de software. A seleção da linguagem adequada é um fator importante para o sucesso de um projeto de BPM (van der Aalst et al. 2003b).

Segundo o *Business Analysis Body of Knowledge* - BABOK (IIBA 2009), um modelo de processo é "uma representação visual do fluxo sequencial e da lógica de controle de um conjunto relacionado de atividades ou ações. A modelagem de processos é utilizada para se obter uma representação gráfica de um processo atual ou futuro dentro de uma organização. Um modelo pode ser utilizado em seu mais alto nível para fornecer um entendimento geral do processo, ou em um nível mais baixo como base para a simulação, para que o processo possa ser tornado o mais eficiente possível".

Para Schedlbauer (2010), a modelagem de processos de negócio inclui o levantamento, documentação, visualização e análise dos procedimentos internos de uma empresa. Segundo o autor, a modelagem de processos é uma arte, exigindo, além de habilidades de engenharia, a capacidade de relacionamento humano. Abordagens tradicionais incluem *brainstorming*, entrevistas, análise de documentos, observação passiva ou ativa e amostragem do trabalho. A aplicação das técnicas de modelagem, embora possa ser academicamente ensinada, possui aplicação pouco precisa.

Vale ressaltar que um modelo de processo não é somente um desenho, exigindo rigor na representação do processo e a adequação a regras semânticas específicas. Um modelo deve representar o processo de maneira concisa e correta, através de elementos visuais e escritos. A construção incorreta de um modelo pode acarretar em interpretações equivocadas e implementações erradas de sistemas, acarretando em um grande retrabalho. Dessa forma, é vital a utilização de uma linguagem padronizada de representação de processos (Schedlbauer 2010).

Durante a modelagem, podem ser identificados conflitos e levantadas discussões acerca da maneira correta de se executar os processos da organização. Tais contradições devem ser vistas positivamente, pois explicitam o grau de desorganização interna atual, e

motivam o planejamento da versão *to-be* dos processos através de um consenso entre todas suas partes interessadas (Valle e Oliveira 2010).

A modelagem tradicional de processos complexos é uma atividade custosa, exigindo uma grande quantidade de tempo e recursos e sendo muitas vezes inviável economicamente (Greco et al. 2006). Como solução para esse problema, diversas técnicas de modelagem que utilizam logs de execução de instâncias de processo têm sido propostas nos últimos anos, com o surgimento do campo de pesquisa de mineração de processos. Dessa forma, é possível obter modelos *as-is*, que explicitam a situação atual dos processos e facilitam sua análise futura. A mineração de processos será estudada em maiores detalhes no capítulo 3 deste trabalho.

2.3.1 Principais notações gráficas

A notação *Event-driven Process Chain* (EPC) foi desenvolvida em 1992 em um projeto de pesquisa da Universidade de Saarland, com participação da SAP AG (Dumas et al. 2005). Ela é suportada pela plataforma ARIS, da IDS Scheer (empresa posteriormente adquirida pela Software AG) . O EPC é utilizado pelo ARIS como integrador e modelador de suas diferentes visões de um processo: organização, dados, controle, função e saída, compondo a arquitetura ARIS (Dumas et al. 2005). Os principais elementos de um diagrama EPC são as funções, que representam as atividades de um processo; eventos, que são condições resultantes de atividades, e que disparam atividades posteriores; e seus conectores (Dumas et al. 2005). Uma vantagem do EPC é que ele permite a conexão de objetos extras, como sistemas e unidades organizacionais.

Tradicionalmente, um modelo bem formado exigiria que dois eventos ou duas funções não se ligassem diretamente, resultando em uma alternância entre estes elementos (Gottschalk et al. 2008). Isso é um ponto negativo do EPC em processos que possuem muitos eventos dispensáveis para o bom entendimento do modelo (Valle e de Oliveira 2009). Recomendações recentes sugerem a inclusão de eventos somente quando existe uma alternância importante entre estados (Baureis 2010) (Valle e de Oliveira 2009), embora a estrutura evento-função-evento seja o padrão da modelagem EPC.

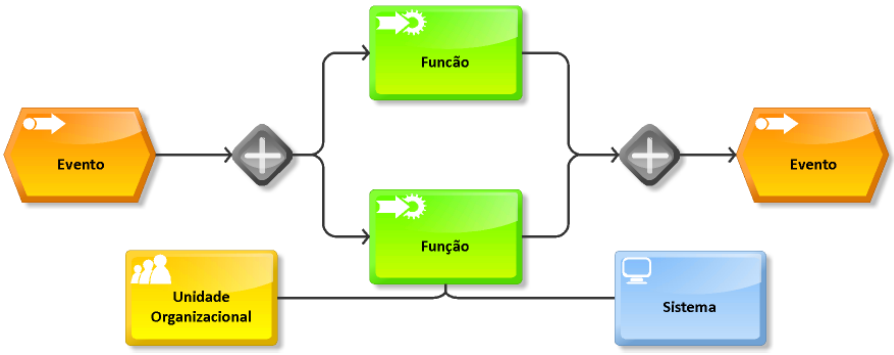

Figura 6 – Diagrama EPC

A *Unified Modeling Language* (UML) é uma linguagem mantida pelo *Object Management Group* (OMG), um consórcio que inclui a participação das maiores empresas da área, como Software AG, IBM, SAP AG, Oracle e CA. Seu objetivo é "auxiliar na especificação, visualização e documentação de modelos de sistemas de software, incluindo sua estrutura e projeto" (OMG 2012). Embora seu foco seja no desenvolvimento de software, a UML permite a modelagem de processos de negócio utilizando seu diagrama de atividades (Valle e de Oliveira 2009). Ao todo, a UML 2.4.1, versão mais recente, possui 14 tipos de diagramas, divididos entre diagramas de estrutura e de comportamento.

O diagrama de atividades UML é de fácil entendimento, e a linguagem é utilizada por uma ampla quantidade de ferramentas, embora estas sejam principalmente voltadas para o desenvolvimento de software. O fato de não ter sido desenvolvida especificamente para a modelagem de processos é uma desvantagem (Valle e de Oliveira 2009), ressaltada pelo fato de a OMG também possuir o BPMN, notação com foco em BPM e com crescente adoção por empresas da área.

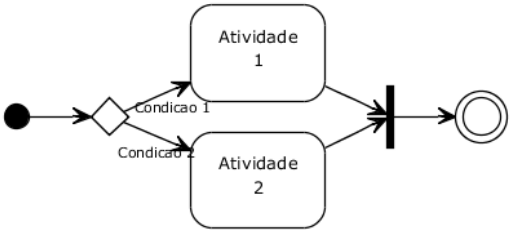

Figura 7 – Diagrama de atividades UML

O *Business Process Model and Notation* (BPMN) (OMG 2011), assim como a UML, é mantido pelo OMG. Ele foi criado com o objetivo de unificar as notações utilizadas pelas diferentes ferramentas de modelagem de processos (Valle e de Oliveira 2009). É possível

converter modelos BPMN para a linguagem *Business Process Execution Language* (BPEL), utilizados por sistemas BPMS. Como existe um esforço entre os desenvolvedores de sistemas BPM para a adoção do BPMN como padrão, e como sua estrutura se assemelha àquela encontrada em logs de execução de sistemas de informação (que geralmente não alternam entre eventos e funções, como é o padrão do modelo EPC), o BPMN foi escolhido como linguagem padrão para este trabalho. A ferramenta desenvolvida para suportar o método MANA implementa um modelador BPMN, que será descrito no capítulo 5. Maiores detalhes sobre os elementos do BPMN serão apresentados na seção seguinte. Detalhes a respeito de outras notações para modelos de processo não descritas neste trabalho, como o IDEF (*Integration Definition*), as Redes de Petri e o YAWL (*Yet Another Workflow Language*), podem ser encontrados em Weske (2007) e Valle e de Oliveira (2009).

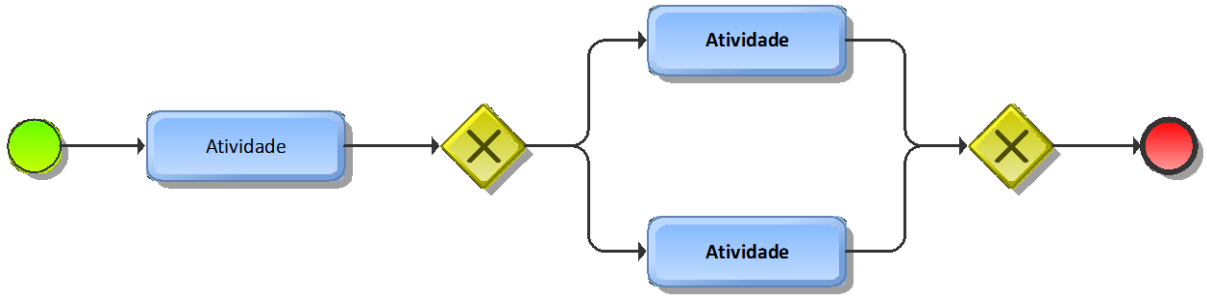

Figura 8 – Diagrama BPMN

2.3.2 *Business Process Model and Notation* (BPMN)

Como dito anteriormente, a notação BPMN (OMG 2011) é mantida pelo consórcio *Object Management Group* (OMG). A versão mais recente do BPMN, a 2.0, foi patrocinada por importantes empresas da área (Earls 2011). O BPMN possui ainda vantagem sobre o EPC para a mineração de processos, pois a estrutura evento-função-evento não é o padrão adotado pelos logs de execução de sistemas estudados. Um dos objetivos do BPMN é facilitar a compreensão dos modelos por todas as partes interessadas do processo, desde os estrategistas e analistas de negócio até os técnicos que implementam tecnologias de gerenciamento de processos (Valle e de Oliveira 2009). A possibilidade de ligar modelos de processo à sua implementação (Valle e de Oliveira 2009) é outro ponto forte, com o mapeamento para linguagens de execução como o BPEL.

Esta seção tem como objetivo apresentar os principais elementos do BPMN. Sua notação define Diagramas de Processo de Negócio (DPN), fluxogramas que utilizam os

elementos descritos pelo padrão (Valle e de Oliveira 2009). Nota-se que a notação possui um número muito maior de elementos do que será apresentado abaixo. O BPMN possui quatro categorias básicas de elementos: *objetos de fluxo, conectores, swimlanes e artefatos* (White 2005). Os objetos de fluxo incluem atividades, eventos e *gateways*.

As atividades representam o esforço de trabalho que será realizado durante um processo (Valle e de Oliveira 2009). Uma atividade pode ser uma tarefa ou um subprocesso, sendo que um subprocesso possui um sinal de mais (+) na parte inferior central da atividade (White 2005).

Figura 9 – Atividade BPMN

Eventos indicam acontecimentos durante a execução de um processo, afetando seu fluxo, e geralmente possuem uma causa e um impacto. Existem três tipos básicos de evento. Os eventos de início indicam quando o processo começa, e são representados por círculos com borda simples. Os eventos intermediários ocorrem no meio de um processo, entre seu início e seu fim, e possuem borda dupla. Os eventos de fim representam o término de um processo, e possuem borda grossa. Os disparadores de um evento (ou seus resultados, no caso de eventos de fim) podem ser indicados por uma imagem complementar no centro do evento (Valle e de Oliveira 2009).

Figura 10 – Exemplos de eventos BPMN

Gateways têm como objetivo controlar convergências e divergências que podem ocorrer durante o fluxo de um processo. Eles possuem a forma de diamantes, utilizada tradicionalmente para estes elementos em diversas notações (White 2005). Os principais tipos de *gateway* são: exclusivo ou XOR, que permite a escolhe de somente uma alternativa de

fluxo a seguir; exclusivo baseado em evento, representando um desvio onde as alternativas dependem de eventos; inclusivo ou OR, que permite a escolha de múltiplos fluxos a seguir; e paralelo ou AND, quanto todos os fluxos de saída são ativados (Valle e de Oliveira 2009).

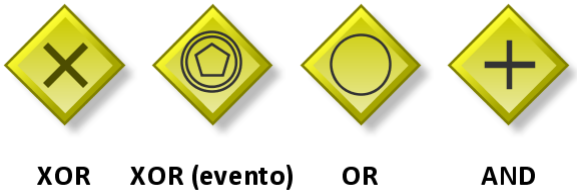

Figura 11 – Principais gateways BPMN

Os conectores são utilizados para conectar os objetos de fluxo, e podem ser de três tipos. Um fluxo de sequência indica a ordem das atividades de um processo. Um fluxo de mensagem indica a troca de mensagens entre diferentes participantes do processo, separados em diferentes *pools*. Finalmente, as associações ligam objetos de fluxo a artefatos, como dados e textos (White 2005).

Figura 12 – Fluxo de sequência BPMN

Swimlanes são utilizadas para organizar atividades, permitindo a visualização de diferentes responsabilidades. Um *pool* encapsula uma unidade de negócio distinta presente no modelo. O conteúdo de um *pool* é considerado um processo autocontido. Dessa forma, somente fluxos de mensagem podem conectar diferentes *pools*, enquanto fluxos de sequência devem conectar objetos de um mesmo *pool* (White 2005). Um *pool* pode ser dividido em diversas raias, separando papéis e funções dentro de um processo. *Pools* representam organizações; raias representam departamentos dentro de uma organização (Valle e de Oliveira 2009).

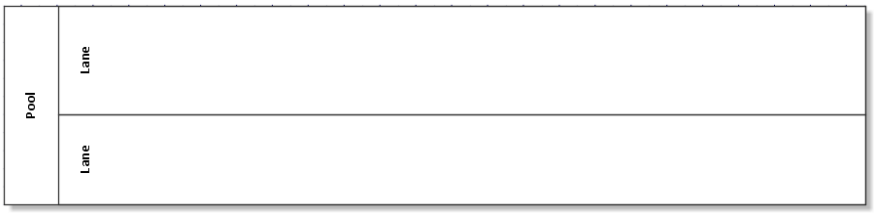

Figura 13 – *Pool* e *swimlanes* BPMN

Artefatos incluem informações adicionais em um modelo BPMN. Eles não alteram a estrutura do processo. Existem três tipos padrão de artefatos pré-definidos: objetos de dados, grupos e anotações. Novos artefatos podem ser incluídos em um modelo de processo, personalizados para a indústria onde o processo é utilizado. Objetos de dados representam dados requeridos ou produzidos por uma atividade. Grupos separam trechos do modelo, para documentação ou análise, sem afetar o fluxo. Anotações fornecem comentários adicionais, utilizados para facilitar o entendimento do modelo (Valle e de Oliveira 2009).

2.4 Análise de Processos de Negócio

A crescente área de *Business Process Analysis* (BPA) inclui os aspectos que não são cobertos por ferramentas de *workflow* tradicionais, como diagnóstico e simulação. No contexto do BPA, ferramentas de *Business Activity Monitoring* (BAM) permitem a utilização de logs de execução de sistemas de informação para a análise de processos. Dessa forma, podem ser extraídas informações sobre fluxo, gargalos, utilização, além da mineração de processos, com a obtenção de modelos de processo a partir de logs de execução (van der Aalst et al. 2003b).

A Gartner (2010) define BPA amplamente como "o espaço de modelagem do negócio onde profissionais do negócio e analistas de TI colaboram na arquitetura, transformação e melhoria do negócio, incluindo a modelagem e a análise de processos para suportar iniciativas de melhoria de processos de negócio", sendo que ferramentas de BPA suportam desde a modelagem de processos ao BAM. Van der Aalst et al. (2003b) também ressalta que a análise de processos tem sentido amplo, incluindo diversas atividades que possuem o objetivo de extrair informações não triviais de processos. Exemplos incluem análise de desempenho, verificação, validação, simulação, dentre outros. Ela permite avaliar o estado atual de um processo, identificar erros de modelagem e propor melhorias e ajustes futuros. A simulação é importante para identificar previamente problemas que possam vir a surgir durante a execução de um processo. A Figura 14 ilustra um exemplo de simulação.

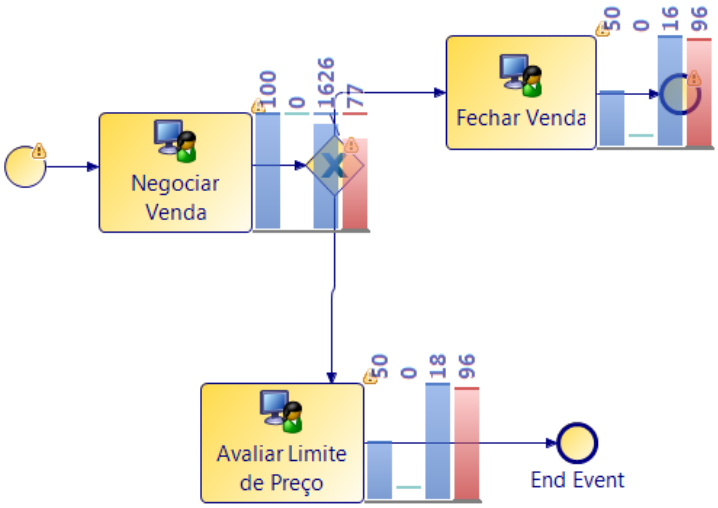

Figura 14 – Exemplo de simulação um processo

Para que a análise de processos ocorra de maneira adequada, deve-se certificar que os modelos de processos existentes para a organização estejam completos e corretos (Schedlbauer 2010). Muitas vezes, uma organização possui modelos defasados e que não correspondem à realidade. Uma maneira de se certificar de que um modelo de processo esteja atualizado é a utilização de algoritmos de conformidade de mineração de processos, que recebem como entrada um modelo de processo e um log de execução desse processo, extraído de um sistema de informação. A mineração de processos será discutida em maiores detalhes no capítulo 3 deste trabalho.

A análise de desempenho representa a identificação de gargalos, envolvendo três dimensões: o tempo de execução do um processo, seu custo e sua qualidade (van der Aalst 2011). Estas análises possibilitam a reengenharia dos processos e sua otimização contínua. Segundo Schedlbauer (2010), a duração e o custo de um processo podem ser calculados somando-se estes valores para cada atividade que o compõe. Para cada valor de duração/custo devem ser feitas três estimativas, de melhor caso (menor tempo/custo para executar o processo), pior caso (maior tempo/custo para executar o processo) e caso esperado (tempo/custo médio para executar o processo). Para identificar gargalos é importante também calcular valores mais detalhados, como, por exemplo, o custo médio de cada atividade ou o tempo médio gasto por cada recurso específico.

Van der Aalst (2011) vai mais fundo, definindo *Key Performance Indicators* - KPIs para cada uma das dimensões definidas. Para o tempo, alguns KPIs incluem o tempo total de execução de um processo; o tempo realmente trabalhado em um processo; o tempo em que um processo ficou esperando pela liberação de recursos; e o tempo em que um processo ficou esperando um trigger para ser liberado, como a finalização de outra atividade. Para o custo, pode-se considerar o tempo médio de utilização de cada recurso. Para a qualidade, pesquisas de satisfação do consumidor, o número médio de defeitos ou o número médio de reclamações seriam KPIs.

A otimização do desempenho de um processo deve ser executada de forma contínua, através de seu monitoramento. A modificação de uma atividade pode ser realizada após uma análise detalhada, identificando, por exemplo, o impacto de ações como sua eliminação, sua simplificação, sua combinação com outra atividade, sua divisão em duas ou mais atividades, sua realocação para outro recurso mais barato e sua automação. A otimização de um fator pode acarretar na redução do desempenho de outro. Por exemplo, a divisão de uma atividade entre recursos mais baratos pode diminuir o custo, mas aumentar a duração do processo (Schedlbauer 2010). Para identificar o melhor curso a seguir é necessário possuir objetivos claros e conhecimento profundo do negócio.

Van der Aalst (2011) ressalta que um grande problema da análise de desempenho tradicional é que ela se baseia em modelos feitos a mão, que muitas vezes são idealizados e não correspondem à realidade. Simulações de desempenho também representam um problema, dado que elas se baseiam em modelos matemáticos, muitas vezes extremamente simplificados. A mineração de processos fornece uma alternativa viável para solucionar estes problemas, pois ela utiliza dados reais de execução de processos, tanto para análise como para gerar modelos realistas para os processos analisados.

O BAM (*Business Activity Monitoring*) é uma subárea do BPA (Gartner 2010) (van der Aalst et al. 2003b), voltada para o monitoramento de processos operacionais em tempo real. Segundo a Gartner (McCoy 2002) o termo é utilizado para definir "como podemos fornecer acesso aos indicadores de desempenho de negócio críticos, em tempo real, para aprimorar a velocidade e a efetividade das operações do negócio". O BAM se diferencia dos demais monitoramentos em tempo real por obter informações a partir de diversas fontes e sistemas, resultando em uma consulta ampla das atividades realizadas dentro da organização.

Para a Gartner (Correia 2002), o BAM se encontra na convergência entre outros mercados, como o *Business Intelligence* (BI) e a integração de aplicativos e *middleware* (AIM).

O foco do BAM é o monitoramento de processos de negócio, utilizando dados em tempo real e históricos coletados de sistemas em operação (webMethods 2006). Ele permite a tomada de decisões orientada a eventos, permitindo que ações sejam disparadas automaticamente a partir de eventos importantes para o negócio (Nesamoney 2004). A correlação de eventos com seu contexto deve ser realizada rapidamente, permitindo, por exemplo, que um gerente seja alertado instantaneamente de uma reclamação feita por um consumidor importante, minimizando o impacto do problema ocorrido (Nesamoney 2004). As ferramentas de BAM calculam medidas de desempenho, ou KPIs, visualizadas através de *dashboards*. Os principais tipos de indicadores de desempenho utilizados incluem volumes (e.g. número de transações, número de eventos), velocidades (e.g. tempo de ciclo do processo), erros e condições especiais definidas pelo usuário (webMethods 2006).

2.5 Business Intelligence

Embora o *Business Intelligence* (BI) não esteja sob o guarda-chuva do BPM, ele foi incluído neste capítulo porque as duas áreas são altamente relacionadas e se unem no chamado *Business Process Intelligence* (BPI). A mineração de processos, parte importante do BPI e o contexto em que este trabalho está inserido, será discutida no próximo capítulo. A Forrester (2010) define amplamente BI como "um conjunto de metodologias, processos, arquiteturas, tecnologias que transformam dados crus em informações úteis e com sentido, utilizadas para permitir compreensões e decisões estratégicas, táticas e operacionais mais efetivas". Porém, a empresa também possui uma definição mais estreita: "um conjunto de metodologias, processos, arquiteturas e tecnologias que alavancam a saída de processos de gerenciamento de informação para a análise, a emissão de relatórios, o gerenciamento de desempenho e a distribuição de informação". Elas se diferenciam no fato de tecnologias de preparação de dados, como ETL e *data warehousing*, ficarem de fora da segunda definição, focada na utilização de dados.

Para van der Aalst (2011), as funcionalidades comuns a ferramentas de BI incluem: ETL (*Extract, Transform and Load*), com a obtenção de dados de fontes diversas, sua transformação para um mesmo formato padronizado e sua carga no armazém de dados;

obtenção de relatórios; buscas *ad-hoc*, permitindo explorar os dados em diferentes granularidades com operações OLAP (*Online Analytical Processing*); *dashboards* interativos; e definição de alertas, disparados a partir de eventos específicos.

Segundo Inmon (2005), BI é a visualização estruturada de dados. Kobielus (2010), analista da Forrester, assume que o termo possui definição nebulosa e de constante redefinição por diferentes especialistas. Van der Aalst (2011) complementa dizendo que diversos termos que entram no guarda-chuva do BI, como *Business Activity Monitoring* (BAM), *Corporate Performance Management* (CPM), e *Business Process Intelligence* (BPI) também possuem definição confusa, reforçada por ações de empresas que buscam distinguir-se das demais através da utilização nomenclaturas e descrições nebulosas.

O *data warehousing* é uma tecnologia central para a exploração de dados realizada no BI. Inmon (2005) define *data warehouse* ou armazém de dados como "uma coleção de dados orientada a assuntos, integrada, não-volátil e variante com o tempo para o suporte a decisões gerenciais". Um armazém de dados contém dados históricos, sendo mantido separadamente dos bancos de dados operacionais. Ao contrário de aplicativos tradicionais OLTP (*On-Line Transaction Processing*), que tratam de operações diárias e repetitivas, com constantes atualizações, os armazéns de dados são voltados para o suporte à decisão, suportando consultas OLAP (*On-Line Analytical Processing*). Seus dados são coletados possivelmente de diversos bancos de dados operacionais, e dados históricos e agregações são priorizados sobre registros individuais. Armazéns de dados são geralmente modelados multi-dimensionalmente. Uma tabela fato, de vendas, por exemplo, contém medidas que podem ser agregadas a partir de diversas dimensões, como data, local, produto e vendedor. Este esquema é chamado de estrela ou floco de neve (quando existe uma hierarquia de dimensões) (Chaudhuri e Dayal 1997).

A carga de dados em um armazém de dados possui alta complexidade, pois seus dados são obtidos a partir de diversas fontes, que muitas vezes possuem dados formatados de maneiras diferentes, redundantes, ausentes ou mesmo conflitantes. Os processos de software que facilitam esta atividade são chamados de ETL, sendo responsáveis por (Vassiliadis 2009): extrair os dados de suas fontes (geralmente bancos de dados relacionais); transportá-los para uma área especializada de processamento do armazém de dados; transformar os dados para se adequarem à estrutura do armazém de dados, além de computar novos valores; limpar tuplas

que não obedecem às regras de negócio; e carregar os dados nas tabelas de destino, atualizando índices e *views* materializadas.

2.5 Considerações finais

Este capítulo introduziu o campo de pesquisa do Gerenciamento de Processos de Negócio, discutindo alguns conceitos da área, que serão utilizados posteriormente, e contextualizando este trabalho dentro do ciclo de vida BPM. Em especial, foi descrita a notação BPMN, que será utilizada para modelar o fluxo de trabalho proposto no capítulo 5 e pela ferramenta desenvolvida nesta dissertação. O próximo capítulo discute sobre a mineração de processos, que está inserida no contexto do BPM. Seu objetivo principal é suportar atividades de BPM utilizando dados reais extraídos de trilhas de auditoria de sistema de informação. Dessa forma, é possível aprimorar as atividades de modelagem e análise em relação às abordagens existentes, que envolvem técnicas como entrevistas e simulação.

Capítulo 3 – Mineração de Processos

Este capítulo disserta sobre a crescente área de mineração de processos, cujo objetivo é utilizar dados reais de instâncias, coletadas em logs de eventos, para auxiliar nas atividades de modelagem e análise de processos. Serão apresentados dois dos principais algoritmos de descoberta de modelos, o algoritmo α e o minerador de heurísticas. O capítulo discute ainda sobre a mineração de processos desestruturados, seus problemas e as abordagens existentes atualmente para lidar com a mineração deste tipo de processo.

3.1 Visão Geral

Técnicas tradicionais de descoberta de processos incluem *brainstorming*, entrevistas, análise de documentos, observação passiva ou ativa e amostragem do trabalho (Schedlbauer 2010). Estas técnicas são manuais, não envolvendo uma análise rigorosa de dados previamente existentes (van der Aalst 2011). Elas demandam uma grande quantidade de recursos e tempo, dos quais uma organização nem sempre pode dispor.

Segundo van der Aalst (2011), a modelagem tradicional é ainda propensa a diversos erros. O analista que projeta um modelo geralmente se concentra no comportamento padrão do processo, deixando de fora, por exemplo, os 20% de casos menos significativos, justamente aqueles que tendem a ser mais problemáticos durante a execução do processo, embora sejam mais raros. Pessoas também podem possuir visões tendenciosas de um processo, variando de acordo com sua função dentro da organização.

Enquanto a maioria dos modelos de simulação de processos se baseia em distribuições de probabilidade fixas, o comportamento humano é difícil de identificar corretamente. Funcionários trabalham em diversos processos concomitantemente, distribuindo sua atenção, o que dificulta a modelagem de um processo individual. Um modelo pode ser criado ainda em um nível de abstração diferente daquele necessário para uma análise específica.

A mineração de processos tem por objetivo resolver os problemas apresentados durante a modelagem e análise de um processo, através da engenharia reversa de dados reais de sua execução, extraídos a partir de sistemas de informação (van der Aalst et al. 2003a). Ela permite avaliar diferentes níveis de abstração de um processo, como, por exemplo, um modelo com 100% dos casos e um modelo com os 80% de casos mais frequentes (van der

Aalst 2011). Ao analisar dados reais de execução, a possibilidade da utilização de visões tendenciosas e generalistas de um processo durante sua modelagem é eliminada. A mineração de processos diminui ainda o custo das atividades de modelagem, permitindo a obtenção mais rápida de modelos, que podem ser refinados posteriormente através de técnicas tradicionais.

A mineração de processos é um campo de pesquisa relativamente recente. O termo surgiu em 1999, a partir de um projeto de pesquisa desenvolvido por Wil van der Aalst e Ton Weijters (van der Aalst 2011). Em van der Aalst e Weijters (2004), os autores definem mineração de processos como "uma metodologia para destilar uma descrição estruturada de processo a partir de um conjunto de execuções reais". A mineração de processos assume que é possível coletar um log de processos contendo a ordem em que os eventos de cada instância são executados. Um log de eventos pode ser obtido a partir de qualquer sistema de informação que armazene estes dados, seja ele um BPMS, um ERP, um CRM ou um sistema desenvolvido localmente pela organização. A partir desta hipótese, diversas técnicas têm sido desenvolvidas e publicadas na literatura.

A tabela 1 exemplifica as principais informações que são extraídas em um log de eventos. Para cada evento do log (linha da tabela), indica-se a instância em que ele ocorreu; a atividade a que ele se relaciona; quem foi o responsável por sua execução; e o momento em que o evento foi registrado. Outras informações podem ser extraídas, como, por exemplo, explicitar se um evento foi de início ou término de uma atividade. Nota-se que o termo *trace*, que indica o registro da execução das atividades de uma instância, é utilizado neste trabalho de maneira intercambiável com o termo *instância*.

Tabela 1 - Exemplo de informações contidas em um log de eventos

Instância (*trace*)	**Atividade**	**Executor**	***Timestamp***
1	Atividade 1	João	01/03/2010
1	Atividade 2	Maria	15/06/2010
1	Atividade 3	Pedro	16/06/2010
2	Atividade 1	Paulo	05/01/2011
2	Atividade 3	Ana	21/04/2011
1	Atividade 4	Rafael	16/04/2011
3	Atividade 1	Ana	08/07/2011
2	Atividade 5	Paulo	10/11/2011
2	Atividade 6	José	02/12/2011

Existem três tipos básicos de mineração de processos (van der Aalst e Gunther 2007):

- *Descoberta*, com o intuito de obter um modelo de processo quando ele não existe previamente. As sequências de atividades de cada instância do processo são analisadas, e um modelo geral é obtido a partir delas. Nota-se que esta análise possui alta complexidade, e os algoritmos de descoberta procuram lidar com diversos desafios como ruído (informações erradas), *overfitting* (quando se exclui casos que não ocorrem diretamente no log), laços, dentre outros. Algoritmos de descoberta serão discutidos em maiores detalhes na seção 3.2.
- *Conformidade*, cujos algoritmos comparam um log de eventos com um modelo pré-existente, analisando se a execução real do processo está de acordo com o que foi modelado. Os desvios encontrados são analisados quanto à sua severidade e sua origem.
- *Extensão*, com o objetivo de enriquecer um modelo com dados do log de eventos, como, por exemplo, projetar informações de desempenho sobre o modelo.

Para Rozinat (2011), a mineração de processos se diferencia do BI pela profundidade da análise realizada. Enquanto as ferramentas de BI se baseiam em indicadores relacionados a processos como um todo, a mineração de processos busca causas nas etapas de um processo, procurando gargalos em dados reais de execução. Armazéns de dados tradicionalmente armazenam somente dados agregados de um processo como um todo, e não de suas etapas. Eles são focados em dados, e desconhecem detalhes dos processos que originaram estes dados (van der Aalst 2011). Van der Aalst vai além, criticando que as ferramentas de BI se focam em *dashboards* bonitos e relatórios simplificados, e não em análises profundas, o que se deveria esperar pelo termo "inteligência".

Castellanos (2009) inclui a mineração de processos dentro do contexto do *Business Process Intelligence* (BPI), que por sua vez, representa a utilização de técnicas de BI para processos de negócio (Grigori et al. 2004). Dessa forma, a mineração de processos estaria dentro do contexto do BI, porém realizando análises aprofundadas e cientes de processo.

3.2 Descoberta de Modelos de Processo

O objetivo desta seção é descrever alguns dos principais algoritmos de descoberta de modelos de processo. Estes algoritmos fazem a engenharia reversa de logs de eventos, que registram a execução das atividades de um processo, em uma abstração de modelo de processo. Um log de eventos contém um conjunto de instâncias ou *traces*, correspondendo a sequências de atividades executadas (Weijters et al. 2006). Um evento registra a execução de uma atividade para uma instância do processo, ou seja, uma *instância de atividade*. Nota-se que, embora a descoberta de um modelo de processo simples pareça trivial, o problema se torna muito mais difícil para modelos maiores. Por exemplo, para um conjunto de 10 tarefas que podem ser executadas em paralelo, o número de possibilidades de traces é 10! ou 3.628.800 (van der Aalst et al. 2004). Além disso, a mineração de processos se diferencia da síntese, que assume como entrada uma descrição completa do comportamento possível. Somente uma fração dos comportamentos possíveis é registrada no log de eventos, tornando inútil um modelo que somente seja capaz de reproduzi-lo (van der Aalst e Gunther 2007).

O restante dessa seção introduz dois dos principais algoritmos de descoberta de modelos, o algoritmo α (van der Aalst et al. 2004) e o minerador de heurísticas (Weijters et al. 2006). O primeiro foi estudado devido à sua importância acadêmica, e motivou o desenvolvimento das técnicas posteriores de descoberta de modelos. O minerador de heurísticas, por sua vez, foi escolhido como a técnica utilizada no restante deste trabalho, devido à sua robustez contra ruído. Outras técnicas de descoberta de modelos e comparações entre elas podem ser encontradas em (van der Aalst et al. 2003a), (van der Aalst et al. 2004), (van der Aalst e Weijters 2004) e (Medeiros et al. 2006). Nota-se que o método desenvolvido neste trabalho não está limitado apenas à utilização do minerador de heurísticas, mas, pelo contrário, qualquer algoritmo pode ser suportado pelo fluxo proposto.

As principais definições utilizadas pelos algoritmos de descoberta de modelos, utilizadas nas seções seguintes, são (Weijters et al. 2006):

- T é um conjunto de atividades. Exemplo: T={A,B,C,D,E}.
- T^* é o conjunto de todas as sequências compostas de zero ou mais atividades de T.
- $\sigma \in T^*$ é um *trace* de eventos, representando uma sequência de atividades. Exemplo: ADEC.

- $W \subseteq T^*$ é um log de eventos, um multiconjunto de *traces* de evento. Sendo um multiconjunto, ele permite a repetição de elementos, ou seja, que uma mesma sequência de atividades ocorra mais de uma vez no log.
- $a >_W b$, para duas atividades a e b, se e somente se existe um trace $\sigma = t_1 t_2 t_3 \ldots t_{n-1}$, e $i \in \{1, \ldots, n-2\}$ tal que $\sigma \in W$ e $t_i = a$ e $t_{i+1} = b$. Ou seja, a atividade b segue diretamente a atividade a em pelo menos um trace do log W.
- $a \not> _W b$ indica que a atividade b não segue diretamente a atividade a nos traces do log W.
- $a \rightarrow_W b$ se e somente se $a >_W b$ e $b \not> _W a$
- $a \#_W b$ se e somente se $a \not> _W b$ e $b \not> _W a$
- $a \| b$ se e somente se $a >_W b$ e $b >_W a$

3.2.1 Algoritmo α

O algoritmo α (van der Aalst et al. 2004) é um dos algoritmos mais estudados na área de mineração de processos, estando implementado na ferramenta ProM (Van Dongen et al. 2005). Seu objetivo é redescobrir uma rede de Petri que modela o processo contido no log de eventos. Mais especificamente, sua saída é uma WF-net, uma rede de Petri que modela o comportamento de um *workflow*. Uma rede de Petri é uma tupla (P,T,F), onde P é um conjunto de posições, T é um conjunto de transições, F é um conjunto de arcos. Em uma WF-net, as transições representam as atividades do processo. As posições e arcos representam as dependências entre atividades. Mais especificamente, as posições são pré-condições ou pós-condições das atividades, permitindo a modelagem de conceitos como os ANDs e ORs entre atividades do processo.

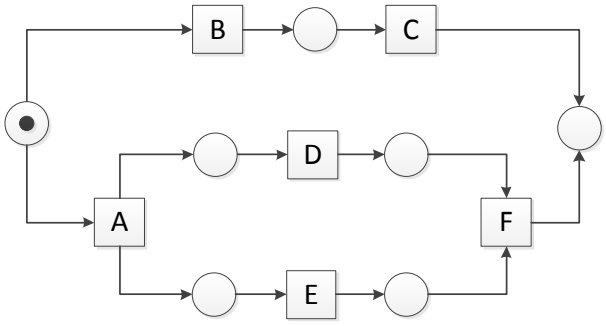

Figura 15 – Exemplo de WF-net - adaptado de van der Aalst et al. (2004)

O algoritmo é definido como (van der Aalst et al. 2004):

1. $T_W = \{t \in T \mid \exists_{\sigma \in W} t \in \sigma\}$,

2. $T_I = \{t \in T \mid \exists_{\sigma \in W} t = first(\sigma)\}$,

3. $T_O = \{t \in T \mid \exists_{\sigma \in W} t = last(\sigma)\}$,

4. $X_W = \{(A,B) \mid A \subseteq T_W \wedge B \subseteq T_W \wedge \forall_{a \in A} \forall_{b \in B} a \rightarrow_W b \wedge \forall_{a1,a2 \in A} a_1 \#_W a_2 \wedge \forall_{b1,b2 \in B} b_1 \#_W b_2\}$,

5. $Y_W = \{(A,B) \in X_W \mid \forall_{(A',B') \in X_W} A \subseteq A' \wedge B \subseteq B' \Rightarrow (A,B) = (A',B')\}$,

6. $P_W = \{p_{(A,B)} \mid (A,B) \in Y_W\} \cup \{i_W, o_W\}$,

7. $F_W = \{(a, p_{(A,B)}) \mid (A,B) \in Y_W \wedge a \in A\} \cup \{(p_{(A,B)}, b) \mid (A,B) \in Y_W \wedge b \in B\} \cup \{(i_W, t) \mid t \in T_I\} \cup \{(t, o_W) \mid t \in T_O\}$, e

8. $\alpha(W) = (P_W, T_W, F_W)$.

O passo 1 cria um conjunto de todas as transições T_W, ou seja, todas as atividades presentes no log. O passo 2 cria o conjunto T_I, contendo as transações de saída da posição de início (tradução livre do inglês *source*), ou seja, todas as atividades iniciais contidas nos *traces* do log. O passo 3 o conjunto T_O, das transações de entrada da posição final (tradução livre do inglês *sink*), ou seja, todas as atividades finais. Os passos 4 e 5 criam conjuntos que auxiliam na descoberta das posições da rede. O passo 4 identifica as posições a partir das relações causais entre as atividades ($a \rightarrow_W b$). O conjunto X_W contém tuplas (A,B), nas quais todas as transações em A são seguidas por todas as transações em B, e nenhuma transação dentro de um dos conjuntos (A ou B) segue outra do mesmo conjunto. No passo 5, Y_W é um refinamento de X_W, eliminando tuplas (A,B) cujos conjuntos estão contidos nos conjuntos de outra tupla (A,B). O passo 6 cria as posições da rede, a partir de Y_W. O passo 7 cria os arcos, conectando as posições e as transições definidas anteriormente. O passo 8 retorna a rede (De Medeiros et al. 2004).

O algoritmo α possui grande importância acadêmica e tem sido alvo de diversos estudos. Muitas extensões para o algoritmo foram propostas na literatura técnica, incluindo, por exemplo, a detecção de laços (De Medeiros et al. 2004) , de atividades duplicadas (Chun-

Qin Gu et al. 2008) (Li et al. 2007) e de dependências implícitas (Wen et al. 2006). Sua abordagem formal, porém, apresenta resultados pouco satisfatórios na maioria das situações do mundo real (Weijters et al. 2006). Isso porque o algoritmo assume que o log não possui ruído, ou seja, que todas suas informações são verdadeiras e relevantes para a modelagem, ignorando a frequência em que cada relação entre atividades aparece. Ele também assume que um log está completo, ou seja, que todas as relações possíveis entre duas atividades estão presentes nele.

3.2.2 Minerador de Heurísticas

O minerador de heurísticas (Weijters et al. 2006) é um algoritmo mais robusto para situações reais, sendo menos sensível a ruído e informações incompletas no log de eventos. Para isso, ele leva em consideração a frequência das relações entre cada par de atividades, com a construção de um grafo de dependência. A probabilidade de existir realmente uma dependência entre duas atividades *a* e *b*, em um log *W*, é dada por $a \Rightarrow_W b$, e definida pela função abaixo. Um valor alto para $a \Rightarrow_W b$ indica que existe uma alta chance de a relação de dependência entre *a* e *b* existir. O número de vezes que a atividade *b* segue diretamente a atividade *a* é representado pela notação $|a >_W b|$.

$$a \Rightarrow_W b = \left(\frac{|a >_W b| - |b >_W a|}{|a >_W b| + |b >_W a| + 1} \right)$$

A tabela 2 exemplifica o cálculo de $\Rightarrow_W$ para o log W={ABD^{13}, ACD^{10}, BC, AD^{2}}. O valor sobrescrito indica quantas vezes *traces* iguais apareceram no log. BC e AD são *traces* incorretos, incluídos como ruído. Utilizando esses valores, o minerador de heurísticas define três limiares configuráveis. Dessa forma, são aceitas no modelo as dependências que:

- Possuem o valor da medida de dependência ($\Rightarrow_W$) acima do *limiar de dependência* (ex. 0.9);
- Possuem frequência de observações acima do *limiar de observações positivas* (ex. 10);
- E cuja diferença entre o valor da medida de dependência da relação e o valor da melhor medida de dependência seja menor do que o *limiar relativo ao melhor caso* (ex. 0,05).

Tabela 2 - Exemplo de aplicação da relação $\Rightarrow_W$

$\Rightarrow_W$	**A**	**B**	**C**	**D**
A	0	0,928	0,909	0,666
B	-0,928	0	0,5	0,928
C	-0,909	-0,5	0	0,909
D	-0,666	-0,928	-0,909	0

Para muitas relações de dependência, porém, o uso de limiares é desnecessário. Isso porque todas as atividades presentes no modelo devem possuir pelo menos uma atividade precedente e uma atividade posterior, dependente da mesma. (excluindo os casos de atividades iniciais e finais). Dessa forma o minerador aplica, além dos limiares apresentados, a *heurística de todas as atividades conectadas*, que escolhe, para cada atividade analisada, sua atividade precedente e sua atividade posterior com o maior valor de $\Rightarrow_W$, incluindo estas conexões no grafo de dependência. Utilizando o exemplo acima, a atividade dependente de A é escolhida entre B (0,928) e C (0,909), sendo que B possui a maior medida de dependência. Para a atividade B, como a atividade A é sua única precedente, esta dependência é novamente escolhida. Para a atividade dependente de B, a relação com D é escolhida. Aplicando esta heurística para o restantes das atividades, o grafo de dependência abaixo é encontrado. Os valores nos nós indicam o número de vezes que a atividade aparece no log, e os valores nos arcos indicam o valor da relação $\Rightarrow_W$. Nota-se que o uso da *heurística de todas as atividades conectadas* é opcional.

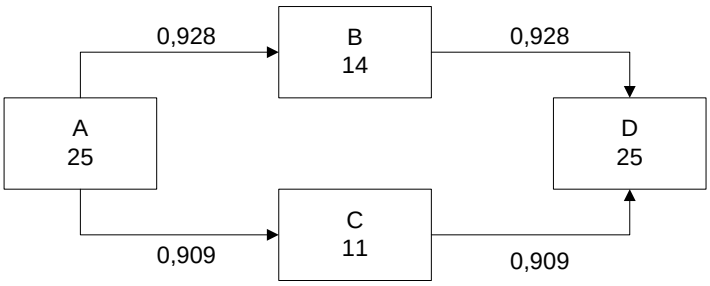

Figura 16 - Grafo de dependência

Um problema da abordagem utilizada até agora é que ela não consegue lidar com laços curtos, ou seja, laços de comprimento 1 (ABBC) e laços de comprimento 2 (ABAC). As medidas de dependência para estes casos retornam valores muito pequenos, o que faz com que os laços sejam eliminados do modelo de dependência resultante. Como solução, o minerador

de heurísticas define novas medidas de dependências para estes laços, e os trata como atividades comuns durante a construção do grafo de dependência. Nas fórmulas abaixo, $a \Rightarrow_W a$ indica um laço de comprimento 1, $a \Rightarrow_{2W} b$ um laço de comprimento 2 e $|a >>_W b|$ o número de vezes em que um laço de comprimento 2 (*aba*) ocorre no log. Ou seja, enquanto a relação $a >_W b$ descrita anteriormente indica que *b* segue diretamente *a* em *W*, a relação $a >>_W b$ representa um laço *aba* existente em *W*.

$$a \Rightarrow_W a = \left(\frac{|a >_W a|}{|a >_W a| + 1} \right)$$

$$a \Rightarrow_{2W} b = \left(\frac{|a >>_W b| + |b >>_W a|}{|a >>_W b| + |b >>_W a| + 1} \right)$$

Outro refinamento do minerador é a identificação de gateways AND e XOR. Considerando-se o grafo de dependência acima, que possui a atividade A conectada por relações de dependência a B e C. Caso A esteja conectada a B e C por um gateway AND, o padrão BC deve aparecer no log. Caso seja por um gateway XOR, o padrão BC não deve aparecer no log. Dessa forma, a medida abaixo é definida. Caso o valor de $a \Rightarrow_W b \wedge c$ seja alto, as atividades estão conectadas por um gateway AND. Caso contrário, elas estão conectadas por um gateway XOR. O corte é feito através de um limiar configurável, como acontece com os demais parâmetros do minerador.

$$a \Rightarrow_W b \wedge c = \left(\frac{|b >_W c| + |c >_W b|}{|a >_W b| + |a >_W c| + 1} \right)$$

O último caso considerado pelo minerador é aquele em que a escolha entre duas atividades não é realizada localmente, mas em outras partes do modelo. A figura abaixo ilustra esta situação. Depois que a atividade D foi ativada, a escolha entre as atividades E e F depende de uma escolha anterior, entre B e C. Enquanto o trace ABDEG é viável, a sequência ABDFG não é. A *heurística de dependência em longa distância* tem por objetivo lidar com esse problema. Ela leva em consideração a relação $a >>>_W b$, indicando a existência da sequência *a...b*, com qualquer número de atividades entre *a* e *b*.

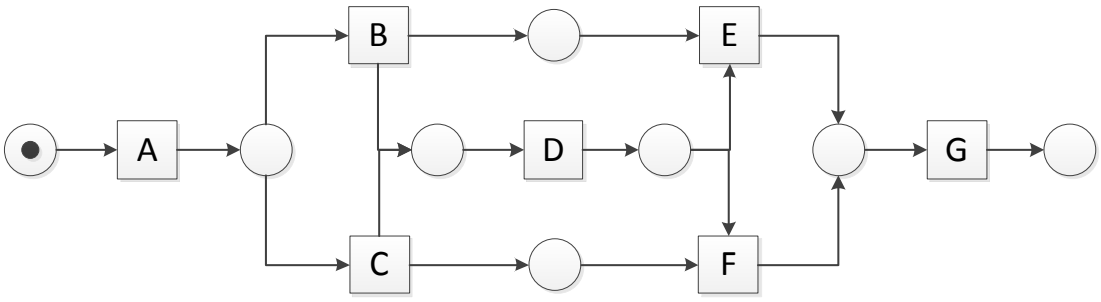

Figura 17 - Grafo de dependência - adaptado de Weijters et al. (2006)

3.3 Processos Desestruturados e Clusterização de Processos

Apesar de seu sucesso na descoberta de modelos de processo a partir de logs de eventos, os algoritmos de mineração de processos apresentados falham em diversas situações do mundo real. Muitos sistemas de informação não amarram a execução de seus processos de uma maneira inflexível, mas possibilitam que eles sejam adaptados caso a caso. Porém, quando é permitido que os usuários de um sistema executem atividades em qualquer ordem, eles irão fazê-lo. O resultado disso são processos desestruturados e muito menos previsíveis do que aqueles executados em um sistema ciente de processo. Ao serem minerados com os algoritmos tradicionais, estes processos geram modelos muito complexos e difíceis de entender. Eles são chamados de modelos de processo em espaguete (Figura 18). A grande quantidade de atividades e arcos torna impossível a utilização prática do modelo para o entendimento do fluxo. Ressalta-se que um modelo em espaguete não está errado, porém o processo que ele tenta descrever é altamente desestruturado (van der Aalst e Gunther 2007).

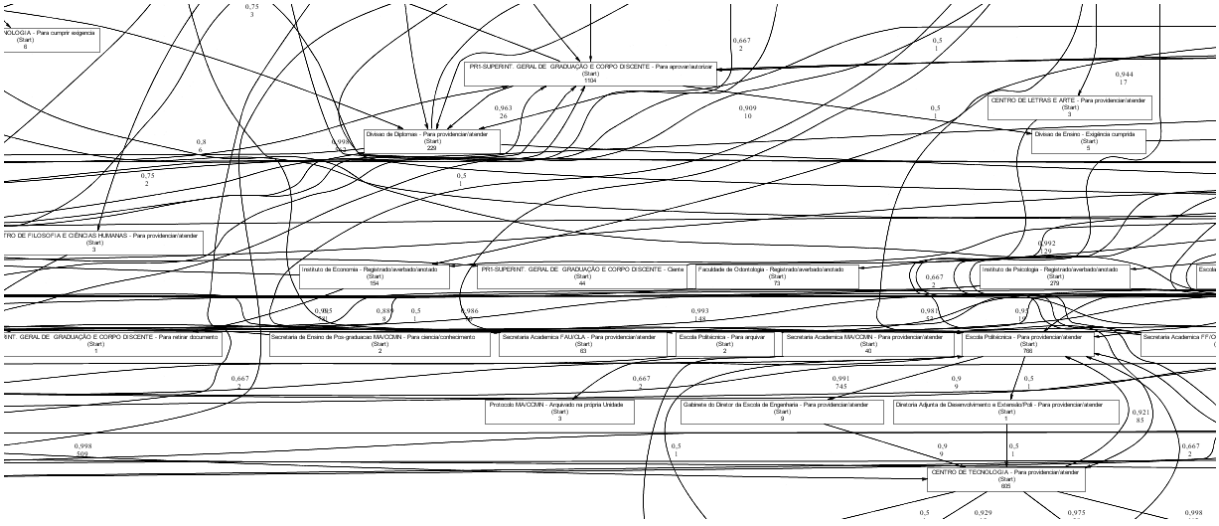

Figura 18 – Trecho de modelo de processo em espaguete

Modelos de processo em espaguete resultam de diversas suposições feitas pelos algoritmos de descoberta, que nem sempre refletem a realidade. Diferentes processos podem estar cadastrados sob o mesmo nome em um sistema de informação (van der Aalst e Gunther 2007). Caso um processo seja tácito, sem modelagem definida, diferentes atores do projeto podem executá-lo de acordo com preferências pessoais. Um processo pode ainda evoluir ao longo do tempo, e diferentes versões de sua execução podem estar representadas em um log de eventos.

Embora existam casos em que as situações descritas acima sejam importantes, como em processos que demandem a adaptação constante de seus atores, muitas vezes uma execução desestruturada reflete uma deficiência da organização. Seu impacto visual pode ser uma grande motivação para o início de projetos formais de reengenharia organizacional e modelagem de processos. Os casos de uso apresentados posteriormente neste trabalho refletem bem esta situação. Sistemas de protocolo de organizações públicas registram cada trâmite realizado em seus processos administrativos. Estes processos são geralmente ineficientes, sem fluxos definidos e mal classificados.

Diversas técnicas têm sido propostas na literatura técnica para lidar com a descoberta de fluxo de processos desestruturados. Em sua maioria, elas procuram quebrar um log de eventos em conjuntos menores de instâncias, através de algoritmos de clusterização. Esta estratégia transforma um problema complexo em vários mais simples. Idealmente, as instâncias de cada cluster devem ser altamente correlacionadas, possuindo fluxos similares. Exemplos incluem o algoritmo de clusterização de *traces* (Song et al. 2008), o *Disjunctive Workflow Schema* (De Medeiros et al. 2007), o algoritmo de clusterização de sequências (Veiga e Ferreira 2010) e a clusterização baseada na distância de edição entre sequências (Bose e van der Aalst 2009). O minerador fuzzy (van der Aalst e Gunther 2007) utiliza outro conceito, agregando grupos de atividades dentro de um mesmo modelo.

As abordagens com impacto sobre este trabalho serão detalhadas nas subseções seguintes. Similarmente aos algoritmos de descoberta de modelos, o método MANA suporta a utilização de qualquer técnica de clusterização. Porém, atualmente a ferramenta desenvolvida suporta apenas o *Disjunctive Workflow Schema* e a clusterização através de um perfil de unidades similar ao algoritmo de clusterização de *traces.*

3.3.1 *Disjunctive Workflow Schema*

A técnica *Disjunctive Workflow Schema* (DWS) (De Medeiros et al. 2007) utiliza uma abordagem de clusterização iterativa para lidar com processos desestruturados. Instâncias identificadas como similares são agrupadas, e um modelo de processo é gerado para cada cluster. O algoritmo de descoberta utilizado é o minerador de heurísticas, devido à sua robustez. Caso um modelo de processo não seja preciso o suficiente, de acordo com os critérios definidos abaixo, ele pode ser particionado novamente, resultando em uma hierarquia de clusters.

O algoritmo busca extrair um conjunto de características relevantes que aparecem no modelo de processo, mas que não estão presentes nas instâncias realmente executadas no log. Elas representam generalizações excessivas no modelo. Estas características são formadas a partir da extensão incremental de cadeias de atividades, até que a cadeia completa seja menos frequente do que suas sub-cadeias. Por exemplo, uma característica relevante seria uma cadeia t_1, t_2, ..., t_n, t_{n+1}, que tenha frequência abaixo de um limiar gama, e que seja composta pela sub-cadeia t_1, t_2, ..., t_n, que seja frequente acima de um limiar sigma, e pela sub-cadeia t_n, t_{n+1}, que também seja frequente acima de sigma. A Figura 19 exibe essa relação de frequência entre uma cadeia e suas partes.

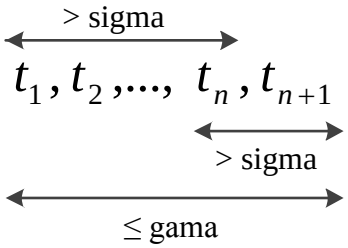

Figura 19 – Identificação de características relevantes pela abordagem DWS - adaptado de Medeiros et al. (2007)

As características relevantes identificadas pelo algoritmo são utilizadas em duas situações. Para clusterizar um conjunto de instâncias, as *m* sequências mais frequentes são utilizadas como atributos para o algoritmo k-means. Uma tabela é construída, contendo uma coluna por característica, e uma linha por *trace* do log. Os valores da tabela, entre 0 e 1, são calculados a partir da porcentagem da sequência analisada que está presente na instância. Por exemplo, caso a característica em questão seja a sequência t_1, t_2, t_3, t_4, e esta sequência completa seja encontrada no trace A, o valor atribuído será 1.

A segunda utilização das características identificadas é como critério de parada: caso um modelo de processo ainda possua generalizações excessivas, ele deve ser particionado novamente para a obtenção de modelos mais precisos. O número máximo de divisões permitidas e o número máximo de clusters gerados a cada divisão podem ser customizados como entrada do algoritmo.

3.3.2 Algoritmo de Clusterização de *Traces*

As abordagens baseadas em clusterização têm em comum a extração de um conjunto de características relevantes de cada instância do processo, que serão utilizadas com algoritmos de clusterização tradicionais da mineração de dados. O algoritmo de clusterização de traces (Song et al. 2008) analisa a similaridade entre *traces* (registros de instâncias) através da construção de perfis. Um perfil agrupa um conjunto de itens relacionados, que representam uma perspectiva das instâncias do processo. Cada item representa uma característica extraída do processo. Para cada item de cada instância é atribuído um valor numérico.

Exemplos de perfis incluem o perfil de atividades, que define um item por atividade encontrada no log, e utiliza como medida o número de vezes que cada atividade aparece no *trace*. A Tabela 3 ilustra um perfil de atividades. O perfil da origem das atividades utiliza uma medida similar, contando o número de vezes que cada executor participou do processo. O perfil de desempenho leva em consideração outros dados, como a duração da instância e a duração mínima, média e máxima das atividades da instância. O perfil de atributos do caso leva em consideração informações adicionais que estiverem anotadas no log, para cada instância. A partir dos dados coletados, a clusterização dos traces pode ser realizada com diversos algoritmos tradicionais de mineração de dados, como o k-means e o *quality threshold*.

Tabela 3 - Exemplo de perfil de atividades

Instância	Atividade A	Atividade B	Atividade C	Atividade D
1	1	2	0	1
2	2	3	0	1
3	0	1	4	2
4	1	2	3	0
5	1	1	2	2

3.3.3 Minerador Fuzzy

O minerador fuzzy (van der Aalst e Gunther 2007) tem como objetivo lidar com dois problemas encontrados em logs reais: os dados extraídos de sistemas de informação não são confiáveis, e nem sempre existe um processo exato que é refletido no log. Esta última característica é a principal causa dos modelos em espaguete. Dessa forma, ferramentas práticas devem suportar o refinamento dos resultados encontrados, através da exploração do usuário, que adquire conhecimento sobre o processo pouco a pouco. A motivação do minerador fuzzy é a mesma utilizada para o método MANA, que permite ao usuário explorar, selecionar e revisar as instâncias de processo selecionadas para a mineração. Ambos, porém, utilizam abordagens diferentes, como será visto posteriormente neste trabalho.

O minerador fuzzy procura estruturar processos desestruturados através de uma abstração comumente utilizada em cartografia. Mapas permitem agregar informação em clusters, combinando, por exemplo, ruas e casas como uma cidade completa. Informações insignificantes para o contexto escolhido são escondidas, e informações importantes são enfatizadas através de dicas visuais como cores e tamanho. Cada mapa é customizado para o seu contexto, dependendo do nível de detalhes exigido, de seu contexto e de seu objetivo.

A técnica utiliza duas métricas: a *significância* das atividades e dos fluxos entre duas atividades, medindo, por exemplo, sua frequência; e a *correlação* entre atividades, medindo, por exemplo, a proximidade entre seus nomes. Com base nessas informações, são realizadas três ações:

- Comportamentos muito significantes são preservados;
- Comportamentos pouco significantes, mas altamente correlacionados são agregados em clusters;
- E comportamentos pouco significantes e pouco correlacionados são removidos do modelo.

A Figura 20 ilustra um modelo de processo gerado a partir do minerador fuzzy, onde octógonos representam clusters de atividades. A figura ilustra ainda o componente de animação presente no minerador fuzzy, que projeta sobre um modelo o andamento das instâncias de processo ao longo do tempo. Essa abordagem de animação de processos foi utilizada como motivação para o módulo de animação desenvolvido neste trabalho.

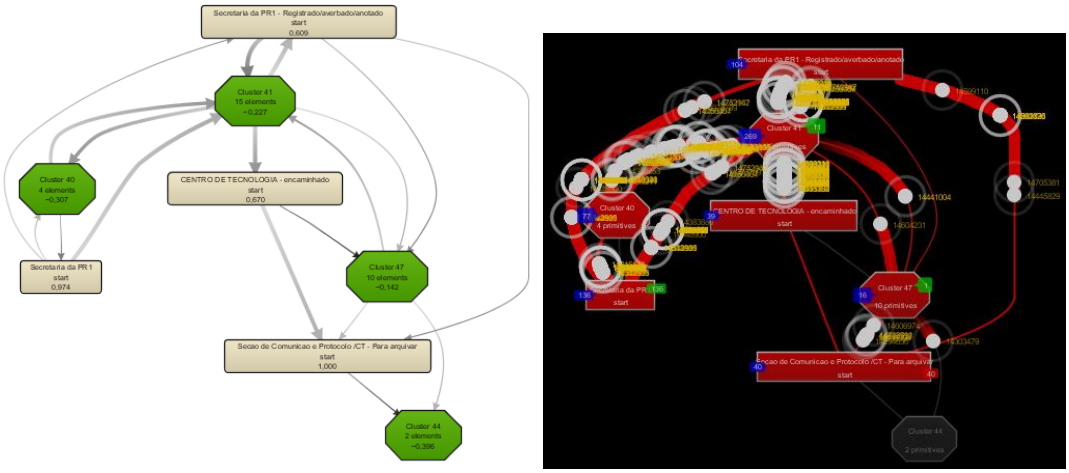

Figura 20 – Modelo gerados com o minerador fuzzy

3.4 Considerações finais

Este capítulo introduziu os principais conceitos da mineração de processos. Alguns algoritmos de descoberta de modelos e de clusterização de processos desestruturados foram apresentados. Em especial, o minerador de heurísticas, o *Disjunctive Workflow Schema* e uma abordagem similar ao algoritmo de clusterização de *traces* foram implementados na ferramenta desenvolvida neste trabalho. O capítulo seguinte tem como objetivo apresentar as ferramentas existentes que suportam a mineração de processos, com foco no framework ProM e no Aris *Process Performance Manager*, que serão comparados com o método MANA no capítulo 5.

Capítulo 4 – Abordagens Similares

Este capítulo introduz as abordagens existentes atualmente que foram desenvolvidas para suportar um fluxo de trabalho de mineração de processos. É dada ênfase ao framework ProM (Van Dongen et al. 2005), detalhado na seção 4.1, pois ele é a ferramenta de mineração de processos mais completa existente atualmente. O ProM foi a única ferramenta encontrada com suporte à mineração de processos desestruturados. Nota-se que este trabalho não tem como objetivo substituir o framework ProM, mas permitir o uso da mineração de processos desestruturados em situações onde as abordagens existentes não retornaram resultados considerados satisfatórios.

O Aris *Process Performance Manager* (IDS Scheer AG 2008), estudado na seção 4.2, permite a análise de processos através de KPIs (*Key Performance Indicators*) e possui funcionalidade simplificada de descoberta de modelos. Essa ferramenta será estudada neste trabalho porque ela utiliza uma base de dados padrão com atributos de instâncias, embora suporte somente processos com tipos e fluxos bem estruturados, permitindo explorar a distribuição de valores de cada uma de suas dimensões.

Van der Aalst (2011) apresenta uma visão geral das demais ferramentas que possuem alguma funcionalidade de mineração de processos. Elas não serão analisadas em maiores detalhes devido ao grande número de ferramentas existentes, à dificuldade de obtenção de versões para teste para algumas delas e ao seu baixo relacionamento com os objetivos deste trabalho, ou seja, a identificação e mineração de processos desestruturados. Exemplos incluem o OKT Process Minig Suite (OKTLAB 2011), que permite ao usuário construir um *workflow* entre um conjunto limitado de atividades de mineração de processos. O Perceptive Reflect (Perceptive Software 2011) fornece funcionalidade de mineração de processos e se integra ao BPMS da mesma empresa. O *Business Process Intelligence* (BPI) (Grigori et al. 2004), assim como o Aris PPM, é baseado em um armazém de dados. Sua funcionalidade de mineração de processos se refere à identificação da origem de comportamentos, derivando automaticamente regras de classificação.

4.1 Framework ProM

O framework ProM (Van Dongen et al. 2005) é uma ferramenta que inclui o estado da arte em técnicas de mineração de processos. Ele funciona através de um sistema de *plug-ins*, permitindo sua extensão com novos algoritmos. Em sua versão 5, estão disponíveis mais de 230 *plug-ins* (da Cruz e Ruiz 2008) que implementam técnicas capazes de minerar diversas perspectivas dos processos, como descoberta de fluxo e conformidade. A Figura 21 resume o funcionamento do framework. Seus componentes principais são (Van Dongen et al. 2005):

- O componente de *filtros de log* importa arquivos de logs de eventos em formato XML, contendo instâncias de processo.
- Os *plug-ins de importação* carregam modelos de processo ou fórmulas para o sistema.
- Os *plug-ins de mineração* realizam a mineração de um processo, utilizando como entrada os logs de eventos. Seu resultado é armazenado em memória e em tela.
- Os *plug-ins de análise* tratam os resultados da mineração e os modelos importados através de análises diversas, como análise de desempenho.
- Os *plug-ins de conversão* transformam resultados em um novo formato, como, por exemplo, uma rede de Petri em um diagrama EPC (*Event-driven Process Chain*).
- Os *plug-ins de exportação* são utilizados para exportar dados, como modelos, para o uso com ferramentas externas.

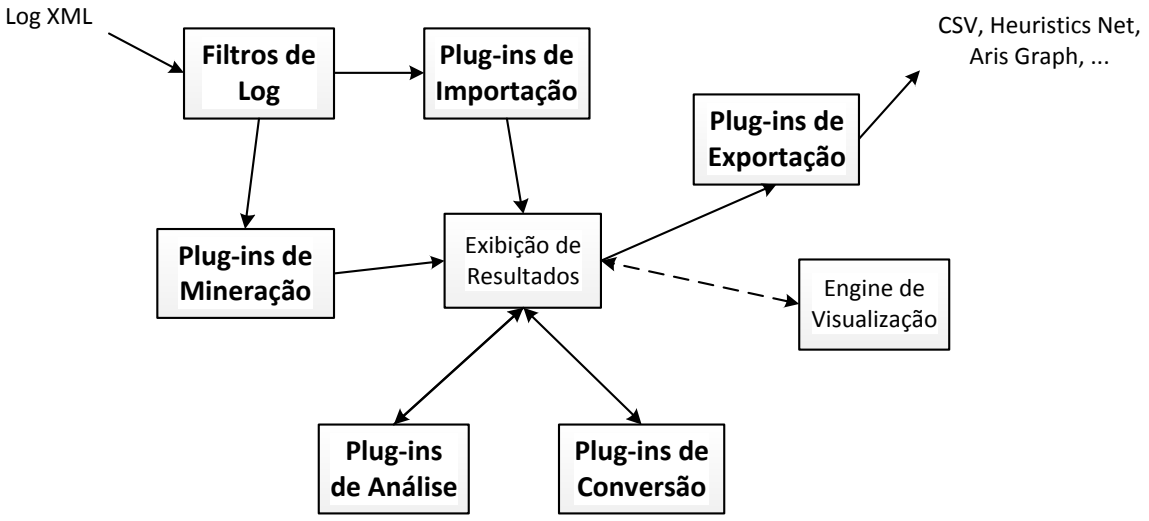

Figura 21 – Estrutura do framework ProM - adaptado de Van Dongen (2005)

Nota-se que o framework não impõe ou sugere nenhum fluxo de trabalho ao usuário. Ao contrário, ele é, na prática, uma coleção flexível de técnicas separadas entre as categorias acima. Além disso, a grande maioria das técnicas exige entendimento prévio para ser utilizada, incluindo muitas vezes a leitura do artigo científico relacionado. Esse fato dificulta sua utilização por usuários que não são especialistas (van der Aalst 2011), sendo uma de suas principais fraquezas, e um dos pontos que o método MANA, desenvolvido neste trabalho, procura tratar.

Duas versões do framework ProM estão disponíveis atualmente: 5.2 e 6.1. A versão 5.2, embora mais antiga, possui uma grande quantidade de plug-ins que ainda não foram portados para a mais recente. A versão 6.1, por sua vez, reformula a interface do sistema, filtrando contextualmente as técnicas disponíveis. Por exemplo, selecionando um log de eventos e uma rede de Petri, somente as técnicas que utilizam estas duas entradas são exibidas. A Figura 22 e a Figura 23 ilustram as duas versões do framework.

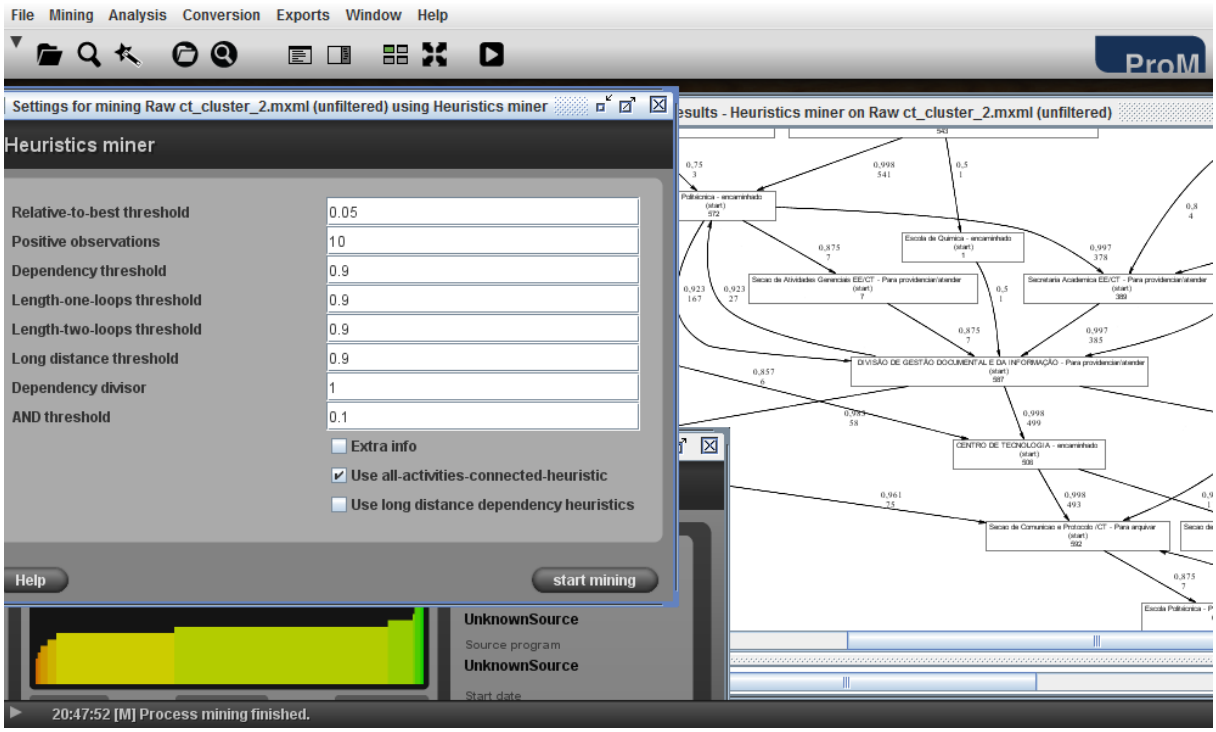

Figura 22 – ProM 5.2

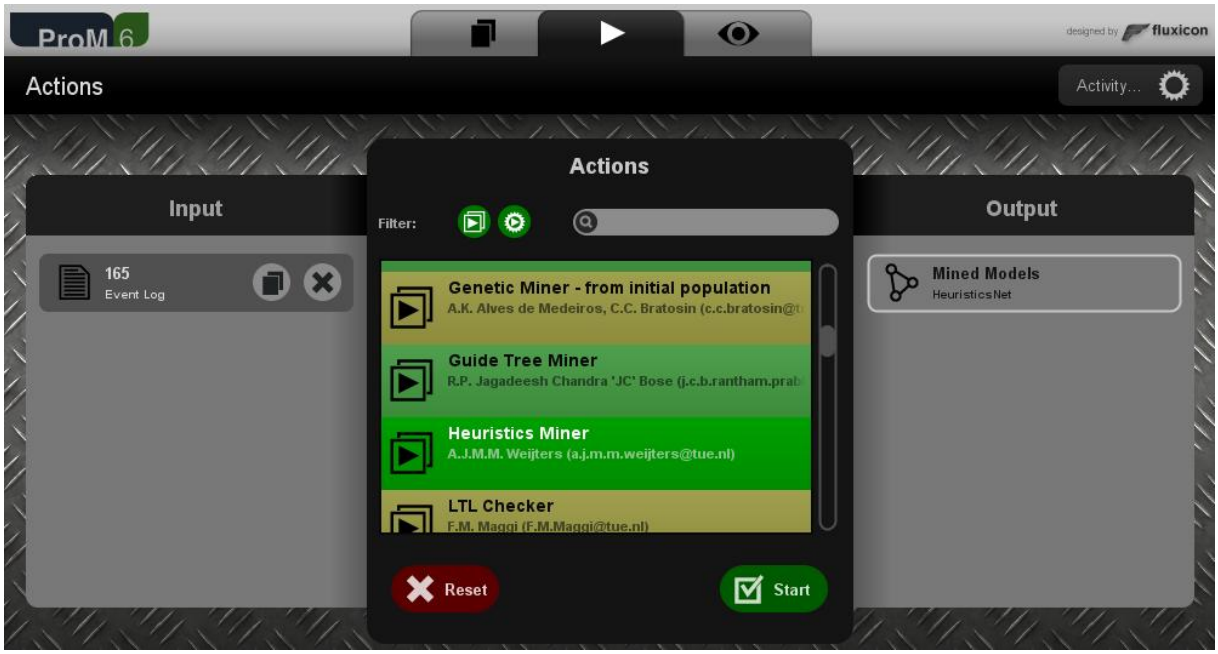

Figura 23 – ProM 6.1

O framework ProM suporta como entrada de dados dois formatos de arquivos XML, para a importação de logs de eventos: o MXML (Dongen e van der Aalst 2005), mais antigo e suportado pelas duas versões, e o XES (Gunther 2009), padrão mais recente e de maior abrangência. O MXML é um modelo bem estabelecido e utilizado pela maioria das publicações na área. Sua estrutura define um log, composto de um ou mais processos. Cada processo é composto de diversas instâncias. Cada instância possui diversas entradas de trilha de auditoria, que correspondem a eventos registrados no log. Uma entrada de trilha de auditoria possui como parâmetros a atividade que foi executada, sua origem, o tipo de evento (atividade se iniciou ou foi completada, por exemplo) e uma *timestamp*. Informações adicionais podem ser inseridas em cada nó da árvore XML (Dongen e van der Aalst 2005)

O XES é um formato mais recente do que o MXML. Ele foi desenvolvido tendo como metas a simplicidade, a flexibilidade, a extensibilidade e a expressividade. Sua estrutura define um log (idealmente de somente um processo), composto de *traces* (instâncias do processo), que por sua vez são compostos de eventos. Cada um desses componentes básicos pode possuir diversos atributos, que são definidos através de extensões ao modelo (Gunther 2009). Dessa forma, o formato busca ser capaz de expressar qualquer caso de uso que possa surgir e que necessite o registro e a troca de informações de um log de eventos.

4.1.1 Extração de Logs de Eventos

Como este trabalho dá grande importância à etapa de identificação de instâncias de processo a minerar, esta seção introduz as ferramentas existentes que extraem arquivos de logs de eventos a partir de trilhas de auditoria. Nota-se que essas ferramentas são voltadas à identificação de quais atributos da base de origem se relacionam a cada atributo do log (e.g. instância, atividade, origem e *timestamp*), e não à exploração dos dados da base e a identificação de instâncias relacionadas, como o método MANA propõe.

O ProM Import (Günther e van der Aalst 2006) é uma ferramenta para transformar trilhas de auditoria de sistemas de informação em arquivos MXML. Os sistemas suportados atualmente incluem SAP R/3, WebSphere Process Choreographer, Staffware, PeopleSoft Financials, CPN tools, CVS, Subversion e Apache 2. O ProM Import pode ser estendido para suportar novos sistemas, através de uma arquitetura de *plug-ins* flexível.

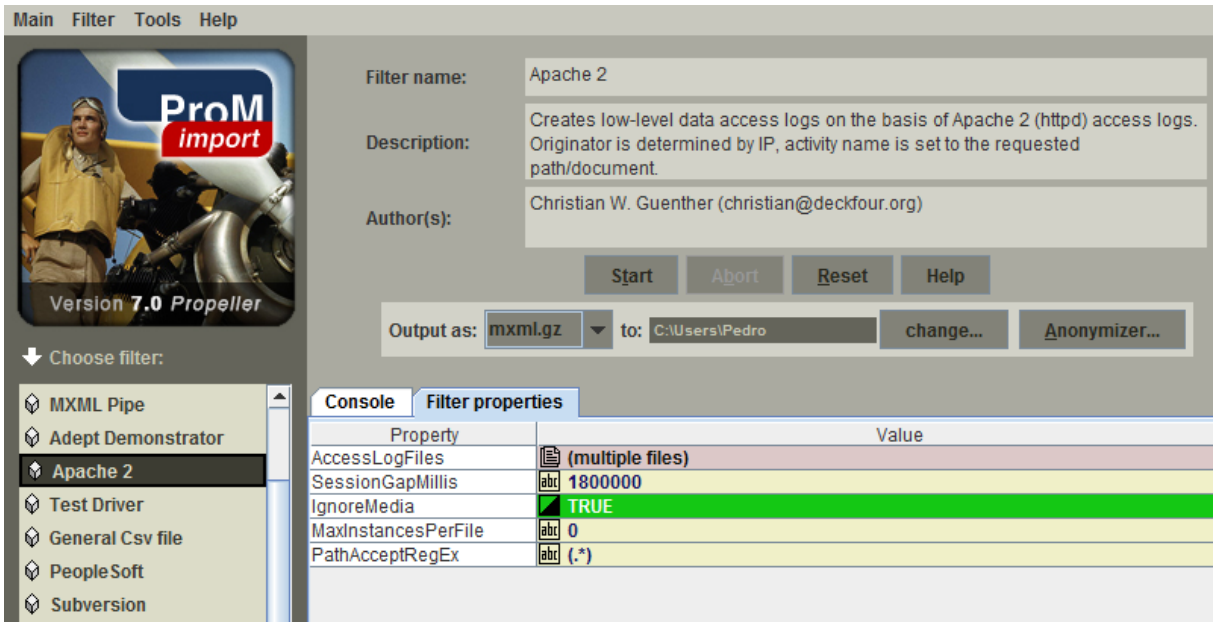

Figura 24 – ProM Import

O XESame (Bujis 2010) é uma ferramenta desenvolvida para facilitar a conversão de bases de dados para o formato XES. Ele funciona através do mapeamento de onde cada atributo do log de eventos desejado se encontra na base de dados de origem, permitindo a conversão de dados por usuários de negócio que não tenham habilidades de programação. Ele é mais abrangente do que o ProM Import, pois não exige a implementação de *plug-ins* para suportar a extração de dados de novas fontes.

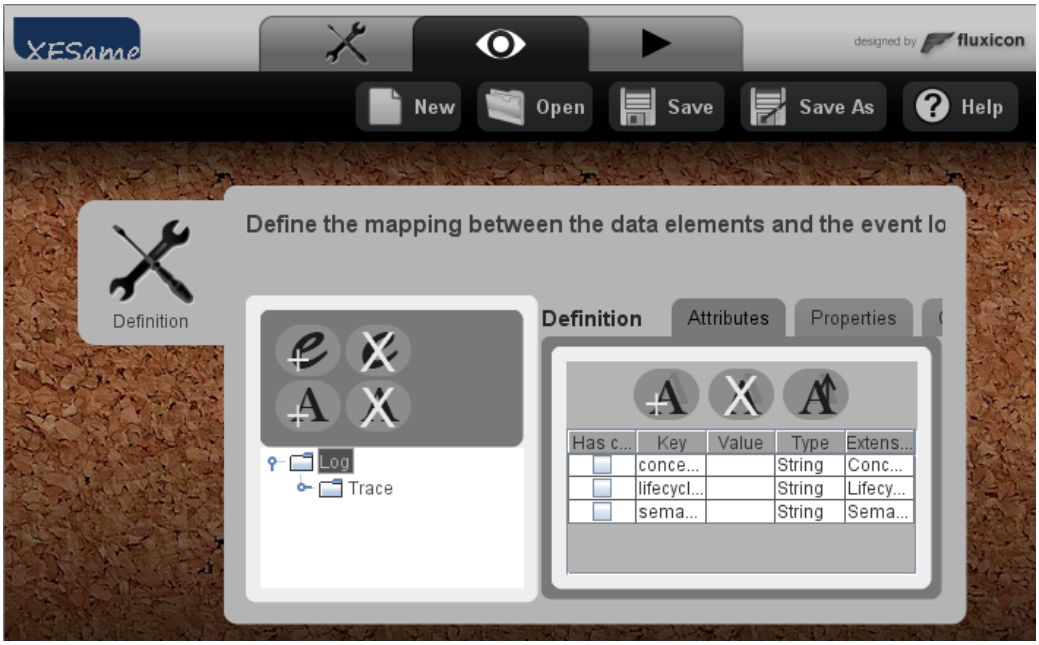

Figura 25 – XESame

O Nitro (Fluxicon 2012) é um sistema que converte arquivos CSV (*Comma-Separated Values*) em arquivos MXML ou XES. A ferramenta permite selecionar as colunas do arquivo importado correspondentes à instância, à atividade executada, ao recurso, à *timestamp* e configurar colunas adicionais. A partir dessas informações, é possível gerar um arquivo log de eventos no formato XES para ser lido pelo framework ProM. Nota-se que o Nitro realiza somente a conversão entre padrões, não suportando a extração de instâncias do sistema de informação de origem.

Figura 26 - Nitro

4.2 Aris *Process Performance Manager*

O Aris *Process Performance Manager* (PPM) (IDS Scheer AG 2008) é uma ferramenta de *Business Process Intelligence* desenvolvida com o objetivo de permitir a análise dos processos de uma organização, suportando diversas perspectivas e indicadores. A ferramenta se baseia em um armazém de dados, para onde são carregados dados da execução de cada atividade de cada instância dos processos que serão avaliados.

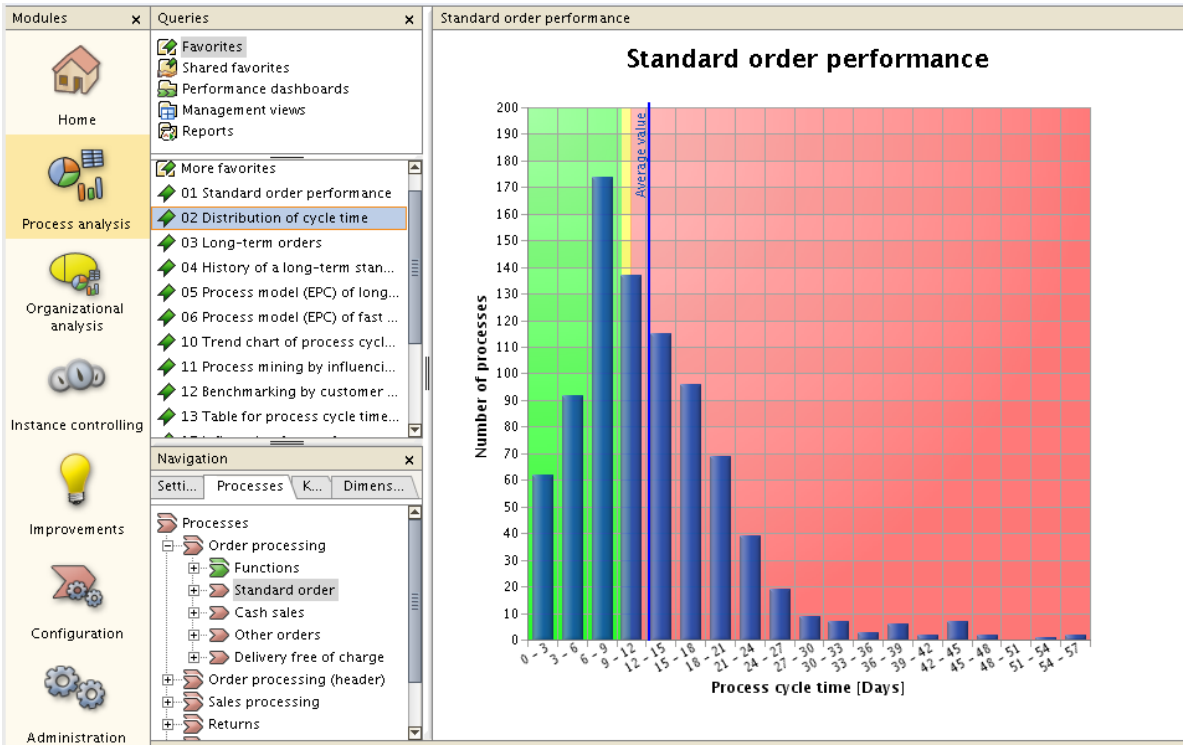

Figura 27 – Aris PPM

O Aris PPM suporta a descoberta de modelos de processo, realçando cada fluxo modelado de acordo com sua frequência detectada, como mostra a Figura 28. Nota-se, porém, que a ferramenta procura reproduzir no modelo todo o comportamento identificado no processo. Isso limita sua utilização em processos contendo ruído, o que ocorre em grande parte das situações do mundo real. O Aris PPM também realiza sua análise separadamente para cada tipo de processo carregado, exigindo alta estruturação dos dados no sistema de informação de origem. A estrutura de processos utilizada pelo sistema pode ser vista no canto inferior esquerdo da Figura 27.

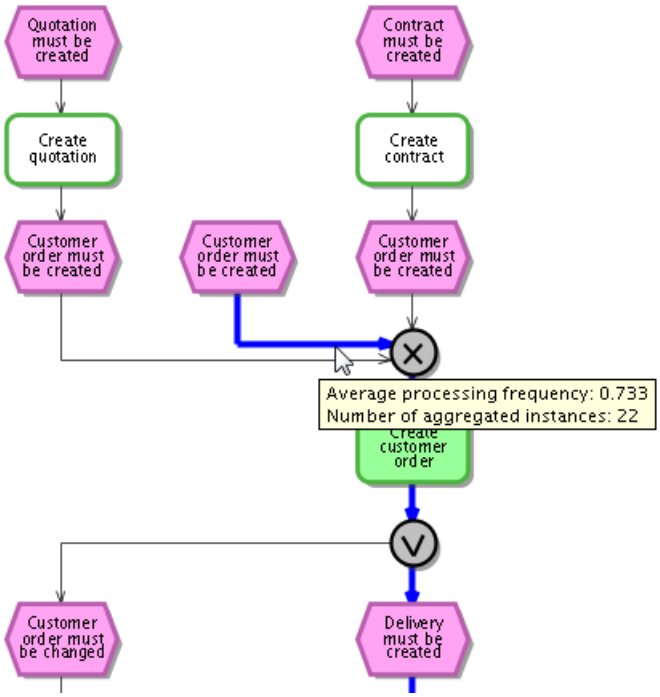

Figura 28 – Trecho de modelo de processo gerado pelo Aris PPM

Os principais conceitos do Aris PPM incluem *tipos de processo, KPIs, valores planejados, dimensões, filtros e consultas* (IDS Scheer AG 2008):

- Todas as instâncias carregadas para o sistema são classificadas em *tipos de processo*, sendo que não é possível comparar tipos de processo diferentes entre si.
- Um *KPI* (*Key Performance Indicator*) representa uma agregação de informações em fatos quantificáveis. Este é um conceito central para o Aris PPM. Um exemplo de KPI seria o tempo total de execução das instâncias. O valor de um KPI é resultado da agregação de todas as instâncias consideradas na análise, como, por exemplo, através de sua média.
- *Valores planejados* podem ser configurados para cada KPI, separando valores limite onde o KPI entra em uma área crítica. Para cada valor planejado pode ser atribuída uma cor. Por exemplo, pode-se configurar um limiar para todas as instâncias cuja duração total seja acima de 2 horas.
- *Dimensões* são utilizadas para analisar e diferenciar KPIs, como tempo, consumidor e vendedor. Dessa forma, o KPI é calculado para cada valor contido na dimensão.

- *Filtros* limitam as instâncias que serão consideradas a partir de limites de valores para os KPIs e as dimensões.
- A utilização da ferramenta para uma análise específica, envolvendo a seleção de KPIs, dimensões, filtros e visualizações, configura uma *consulta*, que pode ser salva para uso posterior. Finalmente, a análise pode ser visualizada através de gráficos ou tabelas. A Figura 29 ilustra a análise de um processo em um gráfico e barras, exibindo o KPI de duração das instâncias, no eixo y, distribuído por país, no eixo x.

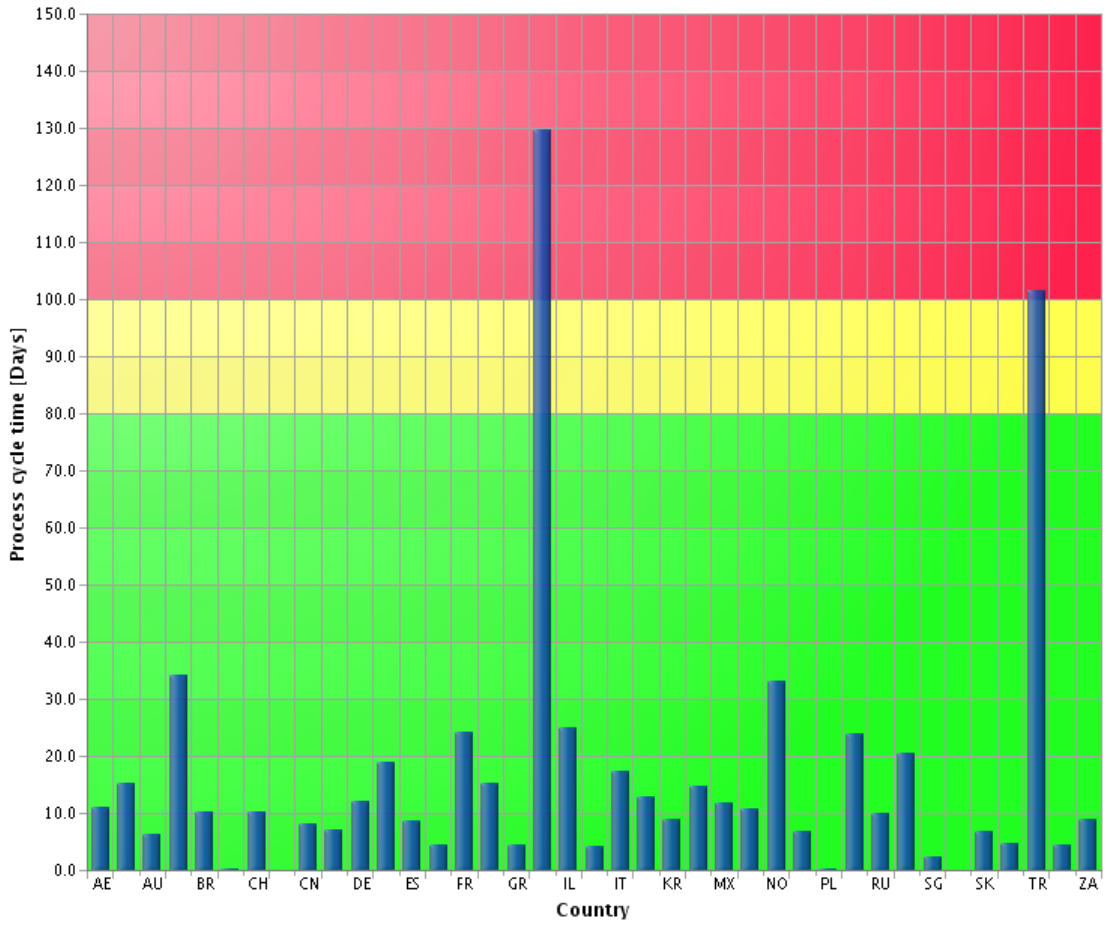

Figura 29 – Duração das instâncias por local de venda no Aris PPM

4.3 Considerações finais

Este capítulo apresentou as duas ferramentas que serão comparadas com a abordagem desenvolvida neste trabalho. O framework ProM é a ferramenta de mineração de processos existente mais completa em termo de funcionalidades. Porém, a dificuldade em identificar grupos de instâncias de processo para carregar na ferramenta e a dificuldade de uso por

analistas não especialistas na área são pontos fracos que este trabalho procura tratar. Por sua vez, o Aris PPM utiliza uma base de dados contendo diversas dimensões dos processos, porém possui apenas funcionalidade básica de descoberta de modelos. Um comparativo dessas abordagens com o método MANA será feito no capítulo 5.

Capítulo 5 – O Método MANA

Este capítulo apresenta o método proposto por este trabalho. A seção 5.1 fornece uma visão geral do método MANA. A seção 5.2 introduz a terminologia utilizada e a modelagem de dados lógica da abordagem. A seção 5.3 apresenta fluxo de trabalho proposto. A seção 5.4 ressalta os principais diferenciais deste trabalho em relação às abordagens existentes. Finalmente, a seção 5.5 detalha os módulos que compõe a ferramenta desenvolvida, e como cada módulo se relaciona à abordagem proposta.

5.1 Visão Geral

Na última década houve uma explosão na área de gerenciamento de processos de negócio. Cada vez mais empresas estão descobrindo a vantagem de se estruturar em torno de seus processos, que devem ser identificados, modelados e gerenciados de maneira correta e eficiente. Em especial, muitas organizações públicas brasileiras poderiam se beneficiar da modelagem formal de processos, pois estes são frequentemente ineficientes, tácitos e desestruturados. O custo, o impacto e o escopo associados a projetos tradicionais de BPM, porém, fornecem empecilhos à sua implantação.

O método MANA foi desenvolvido numa tentativa de alavancar a reengenharia de processos de negócio, especialmente em organizações públicas. Ele procura motivar a importância da modelagem de processos, em ambientes onde sua execução se dá de maneira desestruturada. Para isso, são utilizadas as técnicas de mineração de processos discutidas anteriormente neste trabalho. O principal diferencial do método é que, enquanto as abordagens atuais assumem a existência prévia de tipos de processo bem definidos, esta abordagem utiliza uma base de dados padrão de processos, que podem ser desestruturados, como ponto de partida. Essa base padrão armazena atributos relevantes para a análise de cada instância armazenada. Isso é especialmente importante para dados extraídos de sistemas de protocolo, como os utilizados por organizações públicas para registrar a execução de atividades em seus processos. Esses sistemas muitas vezes não possuem nenhuma classificação entre tipos de processo ou, quando possuem, essa classificação é muito fraca semanticamente, refletindo em uma grande diversidade de processos para cada tipo. Além disso, muitos dados são cadastrados em campos de texto livre, estando sujeitos a erros ou interpretações pessoais. Tais fatores dificultam a extração de logs de eventos para uso em

ferramentas como o ProM. Dessa forma, a atividade de modelagem de processos a partir desses dados deve ser realizada de maneira exploratória.

Em uma busca exploratória, o usuário possui lacunas de conhecimento para navegar em um espaço de informação. Dessa forma, ele submete uma consulta experimental e continua a partir daí, explorando a informação recebida e buscando seletivamente os próximos passos a tomar (White et al. 2006). Uma busca exploratória é motivada "frequentemente por um problema de informação complexo, e um entendimento pobre da terminologia e da estrutura do espaço de informação" (White et al. 2006). Uma busca investigativa envolve a descoberta de novas informações e de gaps de conhecimento, através de múltiplas iterações, por um período de tempo possivelmente longo, sendo que os resultados são analisados criticamente antes de serem incluídos em bases pessoais ou profissionais de conhecimento (Marchionini 2006).

A principal unidade de trabalho do método é a *consulta de processo*. Uma consulta é uma pasta de trabalho, com o objetivo de identificar, através de filtros aplicados sobre a base de dados padrão, um conjunto de instâncias de processo relacionadas. Estas instâncias devem ser posteriormente tratadas por algoritmos de mineração para a obtenção de um modelo de processo. Para suportar o método, foi desenvolvida uma ferramenta de mesmo nome que o apoia na maioria de suas etapas. Ao explorar a base de dados, o usuário pode utilizar filtros, que incluem atualmente: *ano, assunto, descrição, interessado, identificador, origem, situação, unidade inicial, unidade final, unidade participante, atividade inicial, atividade final e atividade executada*. Dessa forma, é possível manter a semântica original da base de origem para a seleção de processos, vital para a realização de uma análise de sucesso. Esta informação seria perdida com a utilização do framework ProM. Caso necessário, uma consulta pode ser clusterizada automaticamente utilizando técnicas discutidas no capítulo 3. O sistema permite ainda a seleção automática de instâncias relacionadas a uma instância específica, utilizando técnicas de busca de vizinhos próximos de raio fixo (tradução livre de *fixed-radius near neighbors*).

O método dá grande importância à análise dos processos analisados a partir de informações visuais. O módulo de animação de processos desenvolvido para a ferramenta permite visualizar o andamento dos processos ao longo do tempo, indicando como as instâncias fluem entre as atividades executadas ou entre suas unidades executoras. A análise

de gargalos, por sua vez, inclui gráficos que permitem identificar intuitivamente as atividades e unidades problemáticas de um processo. O apelo visual das deficiências processuais é importante para motivar tomadores de decisão a alavancarem iniciativas de reengenharia organizacional.

5.2 Terminologia e Modelagem de Dados

Esta seção tem como objetivo apresentar os principais conceitos utilizados pelo método MANA e pela ferramenta desenvolvida, ilustrando seus relacionamentos através de modelos de dados lógicos. A nomenclatura utilizada foi inspirada no padrão XES (Gunther 2009). Para facilitar o entendimento, os conceitos foram divididos em três grupos: *conceitos centrais, conceitos ligados a uma instância de processo, e conceitos ligados a um modelo de processo*.

O termo *processo* é muitas vezes utilizado de maneira ambígua. Dessa forma, o método MANA procura utilizar termos específicos ao se referir a um processo. Um *modelo de processo* é uma abstração do processo, uma "planta" contendo o fluxo que instâncias de um mesmo processo deveriam idealmente seguir. Em uma notação gráfica, um modelo de processo contém um conjunto de nós (atividades, eventos, *gateways*) ligados por arestas. Uma *instância de processo*, por sua vez, é uma única execução deste processo, existindo por um período limitado de tempo. Os mesmos relacionamentos podem ser abstraídos para as *atividades* de um processo, ou seja, *modelos de atividade* e *instâncias de atividade* (Weske 2007). Por exemplo, a emissão de uma única nota fiscal é uma instância de um processo, que idealmente deveria seguir o fluxo definido por um modelo do processo emissão de nota fiscal. Conforme discutido anteriormente, porém, processos nem sempre possuem modelagem explícita, e, mesmo quando possuem, nem sempre seguem o fluxo modelado.

Uma *base de dados padrão de processos* é definida como sendo uma base de dados que registre informações relacionadas a instâncias de processo e das atividades executadas em cada instância. Essa base de dados deve conter um conjunto de dados que auxiliem na identificação de instâncias de processo relacionadas. A base de dados padrão utilizada neste trabalho armazena os seguintes dados, registrados para cada instância de processo: *data de início, assunto, descrição, interessado, identificador, origem e situação*; e os seguintes dados, registrados para cada instância de atividade ou evento: *unidade, atividade e timestamp*. Esses

atributos foram escolhidos pela análise dos dados importantes para o negócio contidos nas bases estudadas no capítulo 6; a base de dados pode ser expandida futuramente para englobar novos casos de uso. A utilização destes atributos para a criação de filtros em cima das instâncias da base será apresentada posteriormente neste trabalho.

Outro conceito utilizado neste trabalho é o de *assunto* ou *tipo de processo*, que são frequentemente descrições genéricas que englobam vários processos. Por exemplo, o tipo de processo *emissão de diploma* de uma universidade, embora à primeira vista pareça ser um processo único, pode possuir, por exemplo, fluxos completamente distintos para diferentes escolas da universidade, ou englobar tanto o registro de novos diplomas quanto a emissão de uma segunda via. Isso é especialmente importante para processos desestruturados, um dos focos deste trabalho.

5.2.1 Conceitos Centrais

- Consulta de Processo: É a principal unidade de trabalho da abordagem. Representa um conjunto de instâncias de processo, extraídas da base de dados padrão a partir de um conjunto de filtros definidos pelo analista de processos. O objetivo de uma consulta é selecionar instâncias relacionadas, que serão mineradas em um modelo de processo. Um modelo gerado a partir da mineração de processos fica logicamente ligado à consulta que deu origem a ele. Pode existir ainda uma hierarquia entre as consultas. A Figura 31 exemplifica um conjunto de consultas de processo organizadas hierarquicamente. Simplificações dos filtros utilizados para definir cada consulta são informadas à direita.

- Tipo de Filtro: Um tipo de filtro representa um atributo contido na base de dados e utilizado para filtrar instâncias de processo. Atualmente a ferramenta desenvolvida suporta os seguintes tipos de filtro: *ano, assunto, descrição, interessado, identificador, origem, situação, unidade inicial, unidade final, unidade participante, atividade inicial, atividade final e atividade executada*.

- Filtro: Um filtro é a utilização de um tipo de filtro em uma consulta, com o objetivo de selecionar as instâncias que fazem parte dela. Os filtros de uma consulta se ligam uns aos outros através de operações booleanas. Um filtro contém um tipo de filtro, um operador de comparação (=, <>, <=, >=) e um valor. Filtros que diferem somente no

valor (com tipo e comparação iguais) são unidos por cláusulas OR. Cada um destes agrupamentos é unido por cláusulas AND, tornando a consulta uma expressão na forma normal conjuntiva. A Figura 31 apresenta alguns filtros. Os caracteres % na figura são uma simplificação para agrupar todas as descrições que contenham o texto informado.

- Instância: Detalhado na seção 5.2.2.
- Modelo: Detalhado na seção 5.2.3.

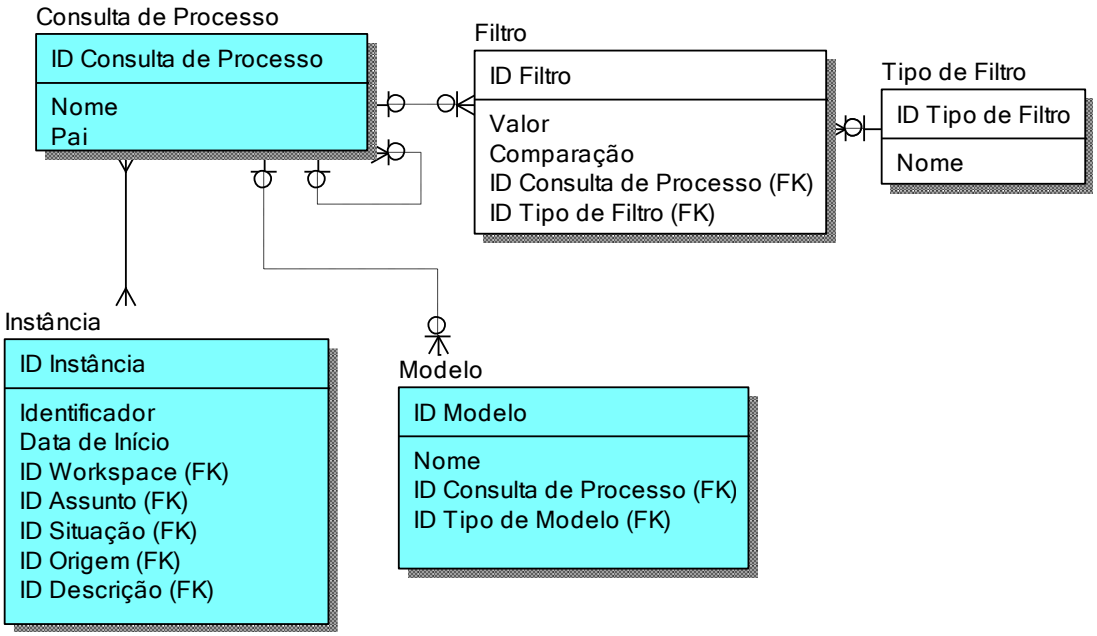

Figura 30 - Conceitos centrais

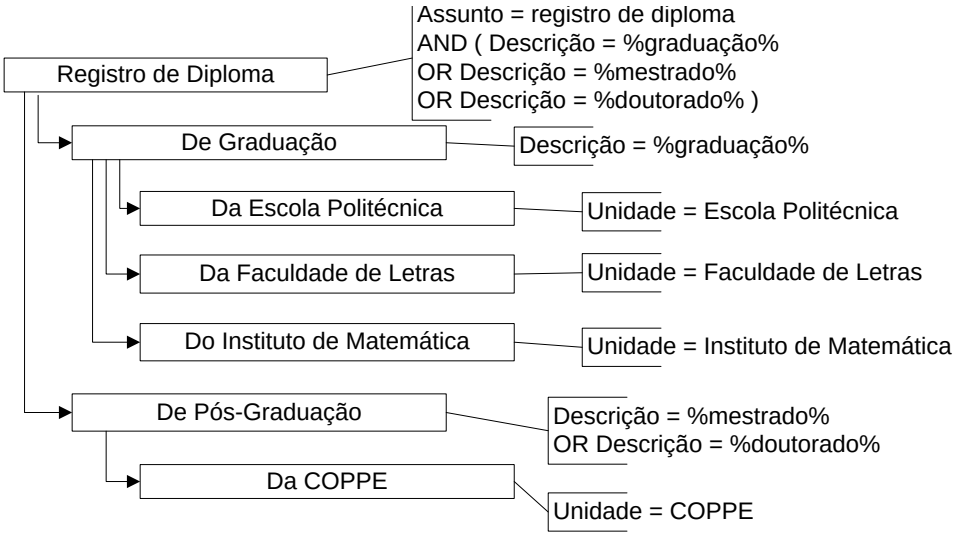

Figura 31 – Exemplo de hierarquia de consultas de processo e seus filtros

5.2.2 Conceitos Ligados a uma Instância de Processo

- Instância de Processo: É uma transação de negócio, uma única execução real de um processo. É comumente usado na literatura técnica de maneira intercambiável com *trace* (quando sua sequência de atividades foi registrada em um log de eventos) ou *caso*. Uma instância existe por um período limitado de tempo, e sua data de início é gravada. Um identificador também é gravado. Uma instância de processo possui um conjunto de instâncias de atividades, registradas em *eventos*, que foram executadas para atingir o objetivo do processo. Assume-se que as atividades são executadas em sequência. Uma instância possui uma origem, um assunto, uma descrição, uma situação (e.g. em andamento, cancelada, concluída), um ou mais interessados e um conjunto de eventos. A Figura 32, centrada na instância de processo, é a representação da base de dados padrão utilizada neste trabalho.

- Assunto: Um assunto é um tipo de processo como identificado na base de dados de origem. Nota-se que um assunto não necessariamente deveria corresponder a um único modelo de processo. Frequentemente, os assuntos preenchidos nas bases de origem são extremamente genéricos ou não são cadastrados. Dessa forma, o assunto de uma instância de processo deve ser utilizado apenas como um guia durante a criação de uma consulta, sendo refinado a partir das demais informações presentes na base, como unidade, descrição, ano e atividades.

- Descrição: Detalhamento de um processo, geralmente com entrada textual do usuário e mais específico que seu assunto. Para situações em que o assunto ligado a uma instância é pouco específico, sua descrição pode ser utilizada para refinar seu entendimento. Utilizando o exemplo da Figura 31, o assunto genérico *registro de diploma* foi detalhado a partir de descrições contendo *graduação*, *mestrado* e *doutorado*. Nota-se que a descrição foi modelada como uma entidade em separado, e não como um atributo da instância. Isso porque, nos casos estudados, as descrições se repetem frequentemente, além de essa abordagem otimizar a execução de filtros sobre este dado.
- Interessado: O interessado é o "cliente" da instância de processo. A existência de interessados é muito comum em processos extraídos de sistemas de protocolo. No exemplo de registro de diploma, o interessado seria o aluno que está se formando.
- Situação: Indica o estado atual de uma instância. Exemplos seriam em andamento, suspensa, cancelada, em cadastramento e concluída.
- Origem: Geralmente indica a unidade responsável pelo processo ou que originou o processo. Nota-se que, nas bases de dados estudadas no capítulo 6, nem sempre as origens são as mesmas unidades que aquelas que participam do processo; dessa forma, as duas entidades foram modeladas separadamente.
- Evento: Um evento registra o início de uma atividade em uma instância de processo. O término da atividade é indicado pela data de início do evento posterior. Um evento grava a atividade executada e a unidade organizacional responsável por sua execução.
- Atividade: É uma unidade de trabalho realizada no contexto da execução de um processo (Weske 2007). Alguns autores utilizam o termo *tarefa* para designar uma atividade. Na modelagem de dados deste trabalho, atividade é uma entidade conceitual que contém seu nome, sendo incluída em outras entidades mais concretas para seu uso. Uma *instância de atividade* está inserida em uma *instância de processo* através de um *evento*. Um *modelo de atividade* está inserido em um *modelo de processo* através de um *elemento de modelo*.

- Unidade: É o executor de uma atividade em um processo, podendo ser uma pessoa, uma organização ou um departamento. O termo unidade foi utilizado porque, nos sistemas estudados no capítulo 6, a execução de tarefas estava delegada principalmente a unidades pertencentes à estrutura organizacional. Em outros tipos de sistema mais detalhados, o executor de uma atividade pode ser, por exemplo, um indivíduo. Na ferramenta desenvolvida, uma unidade pode ser utilizada de maneira intercambiável com as atividades para a descoberta de modelos de processo. Isso porque as atividades dos processos não estavam bem definidas nas bases de dados estudadas, cujo foco se dá no trâmite entre unidades organizacionais. Esta situação será estudada em maiores detalhes nas provas de conceito deste trabalho, no capítulo 6. Outros conceitos relacionados à unidade incluem *unidade inicial*, ou seja, a unidade que executa a primeira atividade registrada para uma instância de processo, *unidade final*, que executa a última atividade registrada, e *unidade participante*, que executa qualquer das atividades registradas para uma instância.

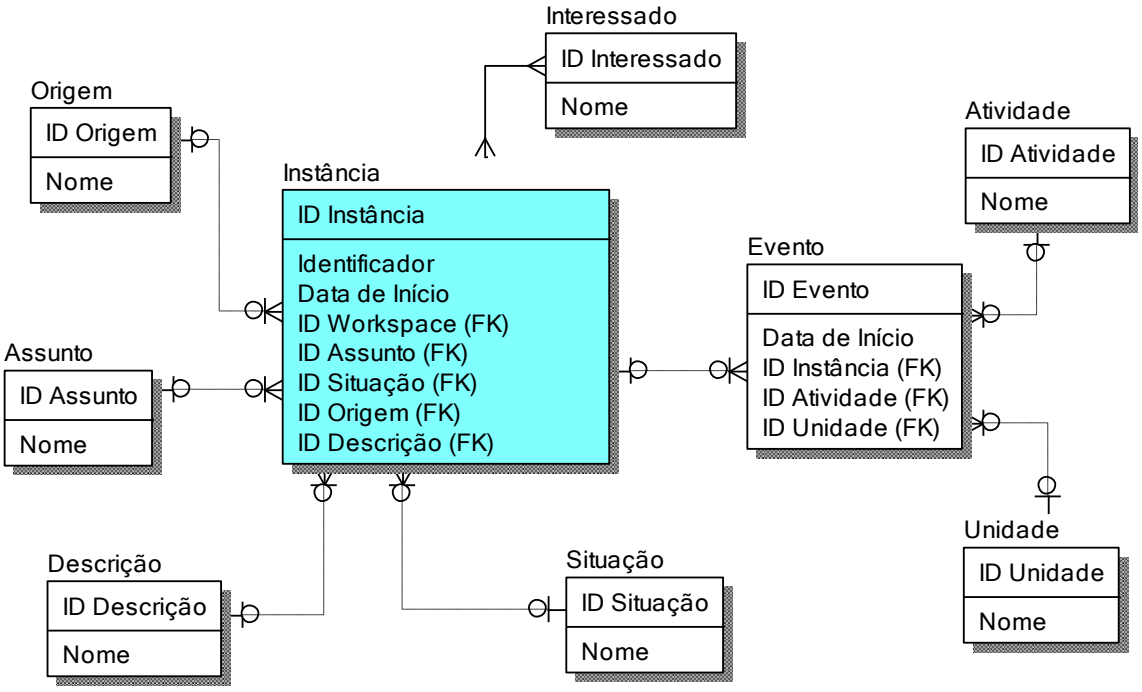

Figura 32 – Conceitos ligados a uma instância de processo

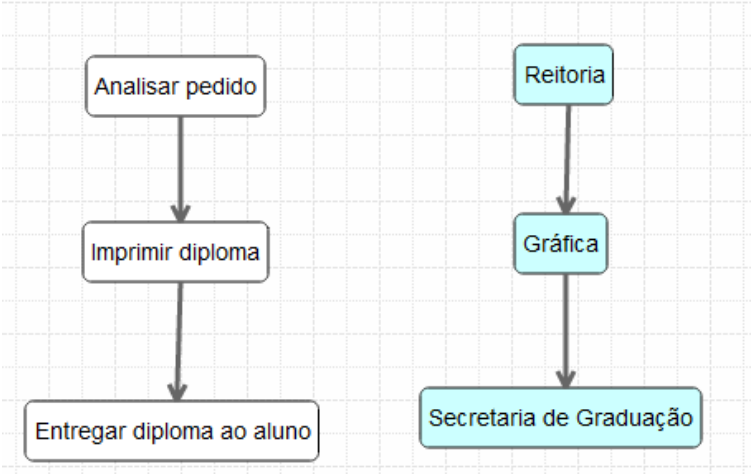

Figura 33 – Exemplos de uma mesma consulta minerada utilizando suas atividades (à esquerda) e suas unidades participantes (à direita)

5.2.3 Conceitos Ligados a um Modelo de Processo

- Modelo: Fluxograma contendo a sequência de atividades que deveria ser seguida durante a execução de um processo. No método MANA, é o resultado da mineração de uma consulta de processo. Um modelo contém diversos elementos, que podem ser atividades, unidades, eventos ou *gateways*.

- Elemento de Modelo: É o componente básico de um modelo de processo. Equivale a um objeto de fluxo de um modelo BPMN. Um elemento é posicionado dentro de um modelo através de uma tupla (x, y).

- Tipo de Elemento: Um elemento de modelo pode ser uma atividade, uma unidade, um *gateway* (AND, OR, XOR, XOR evento) ou um evento (de início, intermediário ou de fim) suportado pela notação BPMN. Nota-se que, embora seus conceitos sejam relacionados, um evento BPMN possui uso distinto daquele dado à entidade evento, descrita acima. A entidade evento marca o início de uma atividade extraída de uma trilha de auditoria, e teve sua nomenclatura herdada do padrão XES. Um evento em um modelo BPMN, por sua vez, é posicionado entre atividades no modelo, não sendo extraído diretamente de uma trilha de auditoria neste trabalho.

- Fluxo: Um fluxo indica uma relação de dependência entre dois elementos do modelo de processo (e.g. duas atividades, um *gateway* e uma atividade). Ele é uma aresta do fluxograma, sendo definido por um elemento de origem e um elemento de destino.

- Atividade: Detalhado na seção 5.2.2.
- Unidade: Detalhado na seção 5.2.2.

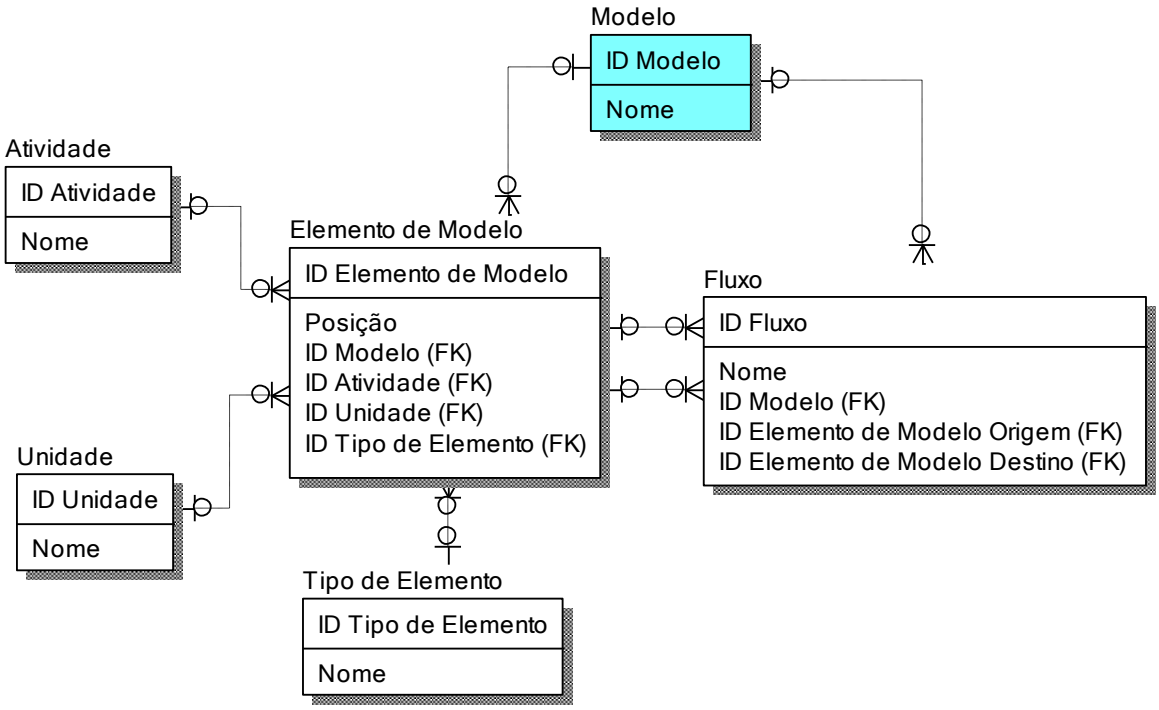

Figura 34 – Conceitos ligados a um modelo de processo

5.3 Detalhamento do Método MANA

Esta seção tem como objetivo detalhar o fluxo de trabalho proposto pelo método MANA. A Figura 35 fornece uma visão geral da abordagem, com o subprocesso principal, de *identificação, mineração, análise e reengenharia de processos*, colapsado. O fluxo se inicia com a carga de instâncias de processo para a base de dados padrão. Isso pode ser feito, por exemplo, através de cargas ETL (*Extract, Transform and Load*). Depois de carregada a base de dados, o processo entra na fase principal de avaliação contínua. Instâncias de processo relacionadas são identificadas; um modelo de processo é gerado através de algoritmos de mineração; o modelo é refinado e validado por suas partes interessadas; as deficiências do processo são identificadas e corrigidas em modelos de processo *to-be*; e o sistema de informação de origem é reavaliado para suportar uma maior ciência de processos, incluindo a melhoria da qualidade de seus dados. A Figura 36, que foi utilizada na introdução deste trabalho, ilustra o relacionamento entre cada fase do método e suas entradas e saídas. As subseções seguintes apresentam detalhes de cada uma das etapas do subprocesso de

identificação, mineração, análise e reengenharia de processos, que é ilustrado pela Figura 37 e pela Figura 38.

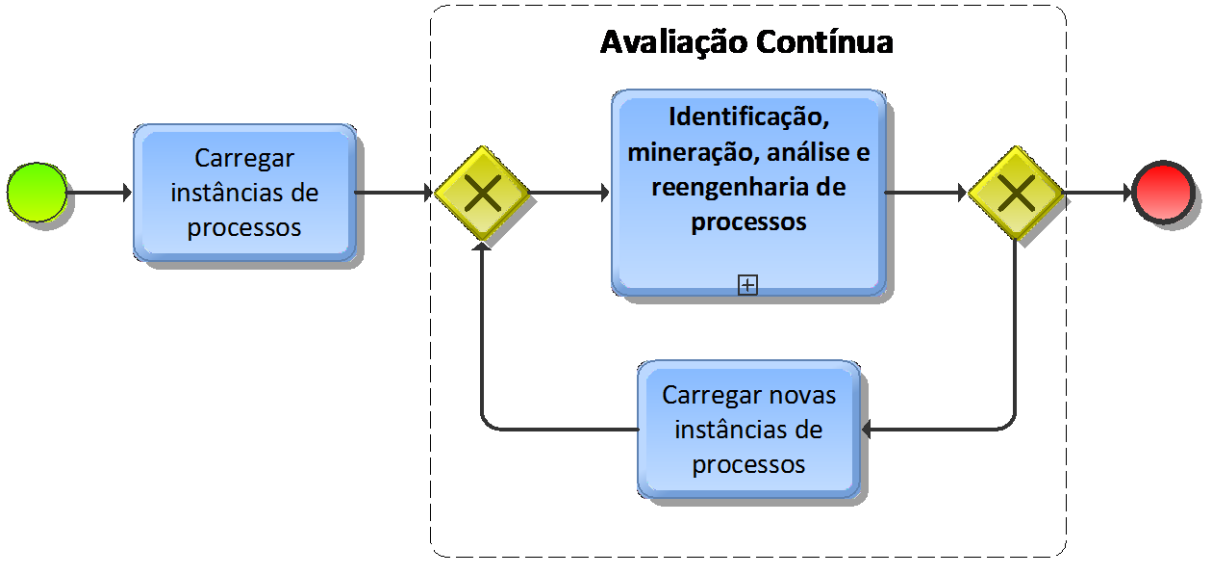

Figura 35 – Fluxo de trabalho do método MANA colapsado

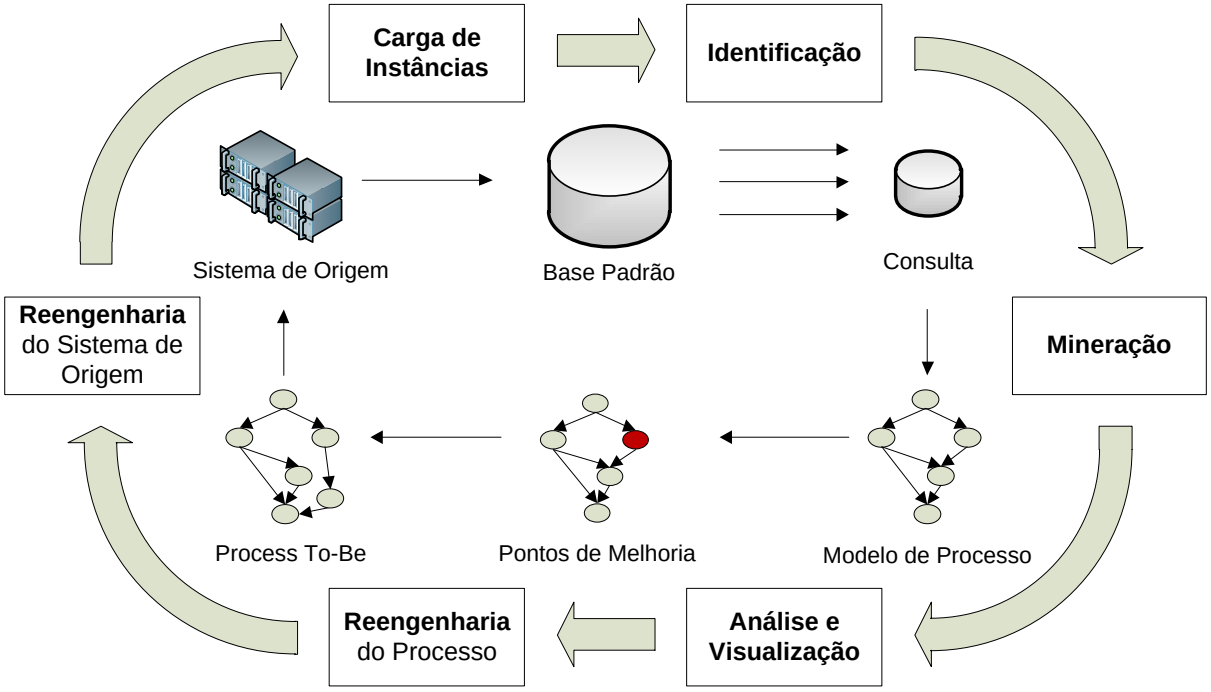

Figura 36 – Entradas e saídas de cada etapa do método

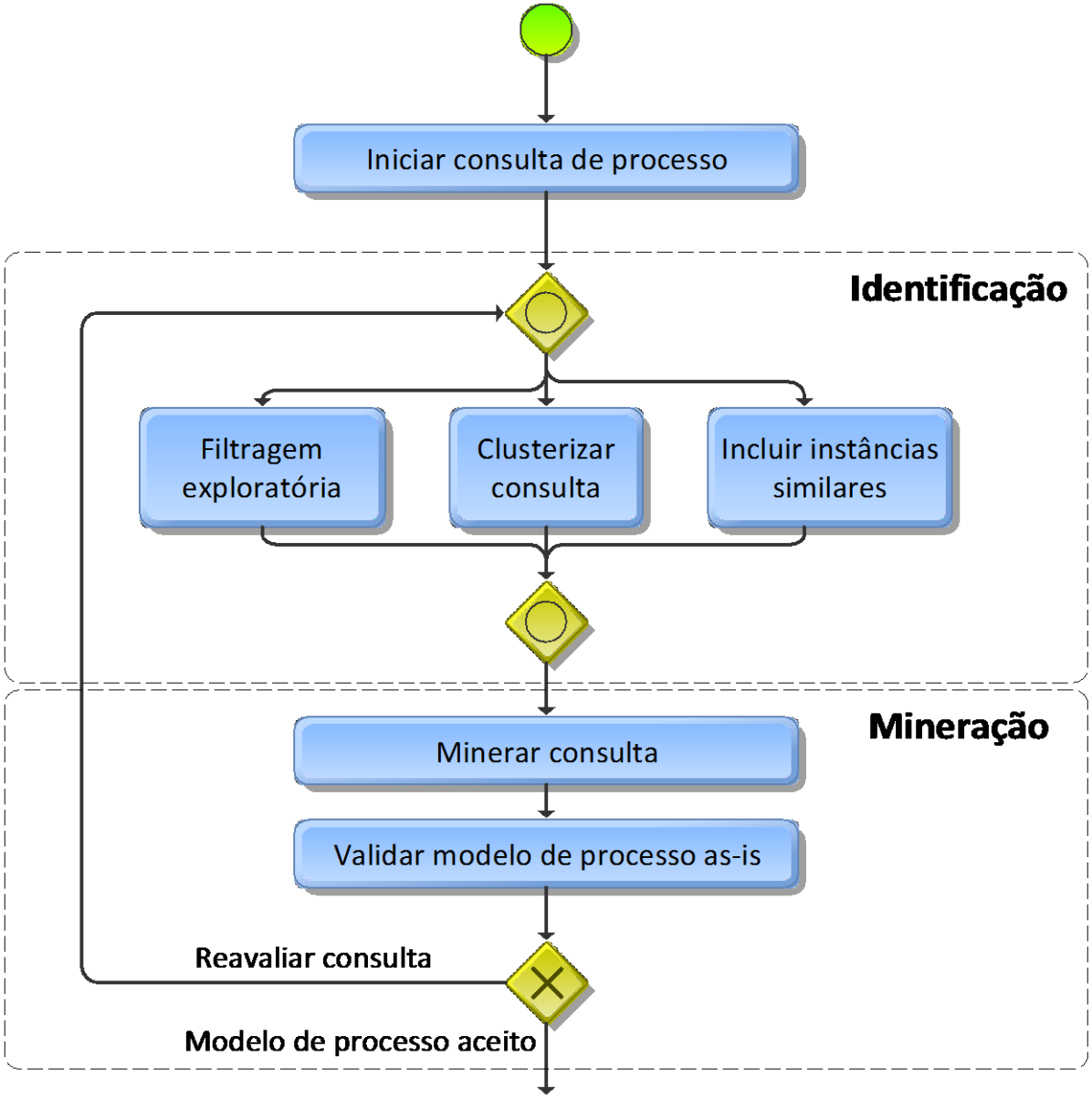

Figura 37 – Detalhamento do processo de identificação, mineração análise e reengenharia de processos, 1ª parte

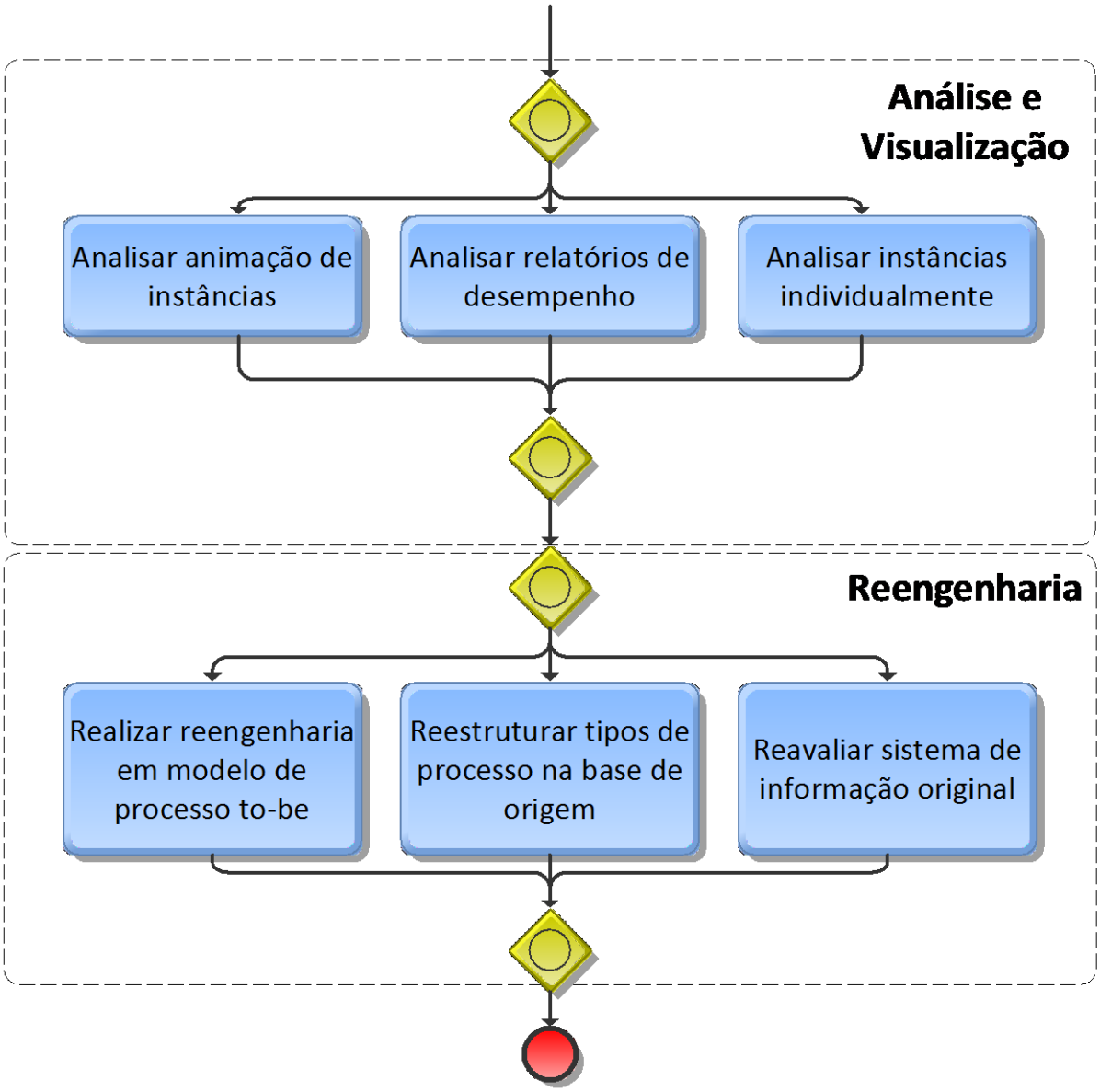

Figura 38 - Detalhamento do processo de identificação, mineração análise e reengenharia de processos, 2ª parte

5.3.1 Identificação

A fase de identificação está relacionada à preparação dos dados que serão utilizados para a mineração de processos, com a seleção de um conjunto de instâncias relacionadas, identificando um processo. O analista inicia o trabalho criando uma nova *consulta de processo*. A consulta é o ponto de partida para a exploração da base de dados padrão, selecionando as instâncias que se deseja minerar. Ela se inicia contendo todas as instâncias da base de dados, sendo que cortes sucessivos são realizados pelo analista, reduzindo o número

de instâncias selecionadas e se aproximando de um conjunto que deveria, idealmente, representar um processo único.

A seleção de instâncias pode ser feita de três maneiras. A principal abordagem proposta pelo método MANA é a exploração dos atributos da base de dados padrão pelo analista de processos. A partir de uma análise inicial da base, o analista pode filtrar as instâncias de processo acordo com suas características. As buscas e filtros suportados são baseados nos atributos chave de um processo da base padrão, como assunto, descrição, origem e unidade participante. Uma busca por todos os assuntos, por exemplo, retornaria todos os assuntos presentes na base de dados, organizados por frequência. Isso permite ter uma visão geral da base. O analista pode então selecionar um assunto de alta frequência como ponto de partida para sua consulta, cadastrando um *filtro*. Em uma universidade, ele poderia optar por trabalhar com processos que possuam *assunto* igual a *registro de diploma*. Buscas textuais também podem ser realizadas. Por exemplo, podem ser pesquisadas as instâncias cujo *resumo* inclua o texto *de graduação*, e incluídos os filtros correspondentes. A filtragem de instâncias em uma consulta é exemplificada pela Figura 39.

Uma consulta pode ser refinada com a inclusão de quantos filtros forem necessários. O filtro de *unidades participantes*, por exemplo, permite analisá-las de acordo com sua frequência, possibilitando a remoção de casos pouco frequentes da consulta. Quando uma unidade somente participou de uma instância, esse caso tem alta probabilidade de ser um *outlier*. O analista pode escolher ainda trabalhar somente com instâncias que passaram, por exemplo, pelo departamento de TI. O filtro de *situação* permite remover as instâncias que estejam incompletas. A análise por *ano* permite selecionar somente um período de tempo desejado para a mineração do processo. Nota-se que é possível criar uma hierarquia de consultas. O exemplo da Figura 39 mostra uma consulta de *contratações da TI ou do RH*, que possui as subconsultas *contratações da TI ou do RH em 2010* e *contratações da TI ou do RH em 2011*.

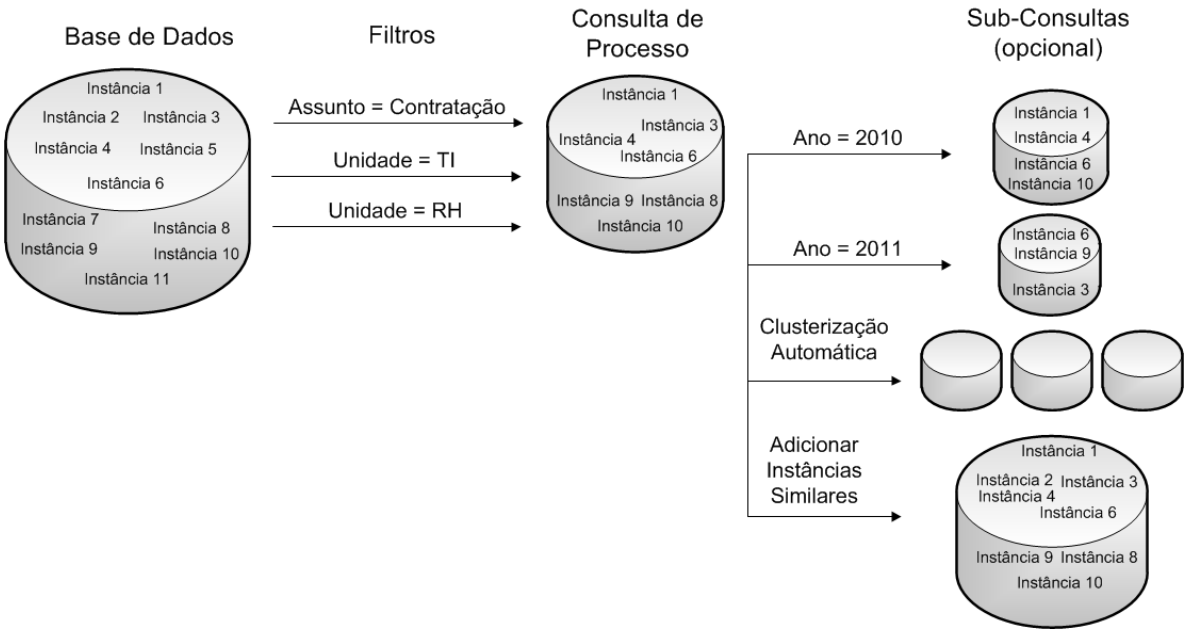

Figura 39 – Filtragem e hierarquização de consultas

Vale ressaltar que a busca exploratória a partir dos atributos de uma base padrão está relacionada a abordagens de busca facetada (tradução livre de *faceted search*). Esse tipo de busca envolve a classificação de um item em categorias ortogonais, permitindo que o usuário refine sucessivamente sua busca a partir dos valores disponíveis para cada uma delas (Yee et al. 2003). Seu uso é muito frequente na Web, principalmente em sítios de comércio eletrônico (Ben-Yitzhak et al. 2008). A busca facetada, porém, está geralmente associada à recuperação de itens individuais de uma coleção, auxiliada pelas categorias existentes. A abordagem de busca deste trabalho, por sua vez, tem como objetivo a obtenção de um cluster de itens. Dessa forma, é dada maior importância aos valores presentes em cada categoria do que à exibição individual de itens (instâncias de processo).

Existem casos em que mesmo a identificação exploratória de instâncias relacionadas resulta em processos em espaguete. Para isso, devem ser utilizadas técnicas automatizadas. Uma abordagem possível seria utilizar algoritmos de clusterização, introduzidos no capítulo 3 deste trabalho. Isso é importante quando um processo possui fluxos de atividades distintos, que podem ser separados, mas que não estão explícitos ao analista de negócio. Por exemplo, o processo de *registro de diploma*, para os formandos do curso de Engenharia da Computação de uma universidade, pode seguir um fluxo distinto em relação ao processo do curso de Engenharia Civil. Esta situação pode não ter sido prevista pelo analista, ou podem não existir

atributos suficientes na base de dados para separar manualmente estes processos. Os clusters resultantes desta etapa podem ser novamente filtrados, para uma nova remoção de *outliers*.

Outra abordagem automatizada seria a identificação de instâncias similares. Isso é feito através de técnicas de busca de vizinhos próximos de raio fixo (tradução livre de *fixed-radius near neighbors*). Essa abordagem se inicia a partir da identificação de uma instância que seja relacionada ao processo que se deseja analisar. A partir dessa instância são identificadas as demais instâncias da base padrão que possuem características similares a ela. A ferramenta desenvolvida neste trabalho implementa uma abordagem para isso, utilizando o *fingerprint* das descrições das instâncias. Essa abordagem será discutida na seção 5.5.

5.3.2 Mineração

Uma *consulta* representa uma seleção de instâncias de processo relacionadas, que foram definidas por um conjunto de filtros. O objetivo da etapa de mineração é obter um modelo de processo a partir destas instâncias, utilizando os algoritmos descritos no capítulo 3. Os algoritmos de mineração permitem que alguns parâmetros sejam controlados, refinando o resultado obtido. O modelo pode conter, por exemplo, todos os comportamentos presentes no processo, ou somente seu fluxo principal. O resultado da mineração é um modelo de processo *as-is*, ou seja, contém o fluxo atual de execução do processo.

Figura 40 – Mineração de processos a partir de uma consulta

O modelo minerado deve passar posteriormente por uma etapa de validação e refinamento, tanto pelo analista quanto pelas partes interessadas pelo processo. Reuniões devem ser executadas para discutir o modelo resultante da mineração e enriquecê-lo com novas informações quando necessário. Pode ser preciso ainda realizar alterações cosméticas de *layout* no processo. O impacto causado pela análise de modelos *as-is* é de extrema

importância para motivar projetos de reengenharia de processos, visto que mostra o grau de desestruturação dos processos da organização.

5.3.3 Análise e Visualização

A fase de análise está relacionada ao diagnóstico do processo minerado, com a identificação de pontos passíveis de melhoria. O método MANA propõe inicialmente três tipos de análise, focando-se na visualização de dados. Como as técnicas apresentadas abaixo estão fortemente ligadas às suas implementações, elas ficarão mais claras na seção 5.5, durante a apresentação da ferramenta desenvolvida para suportar a abordagem. Como é dada ênfase à análise utilizando informações visuais, o termo *visualização* foi incluído no nome desta fase.

Uma abordagem importante é a animação de um modelo de processo, projetando sobre ele o andamento das instâncias que geraram o modelo. Dessa forma, é possível visualizar intuitivamente o desempenho do processo ao longo do tempo e identificar atividades ou unidades que mereçam ser uma avaliação mais detalhada durante a construção de um modelo *to-be*. O impacto visual causado pela animação de processos é importante para motivar ações de reengenharia de processos para níveis hierárquicos mais altos das organizações.

A utilização de relatórios gerenciais para a análise de desempenho também é importante na identificação de sinais de lentidão nas atividades ou unidades de cada processo. O método MANA prevê a emissão de relatórios para informar os tempos mínimo, médio e máximo que cada atividade levou para ser executada. Dados que foram agregados em cada linha dos relatórios (e.g. agregando todas as execuções de uma mesma atividade em uma linha) podem ser visualizados separadamente utilizando gráficos, como *scatter plots* e linhas do tempo.

A identificação de quais instâncias do processo foram executadas de forma ineficiente é importante para o rastreamento das causas dos problemas identificados. As análises descritas acima, utilizando técnicas de animação, relatórios e gráficos, devem permitir ao analista de processos destacar uma única instância que mereça ser analisada em maior profundidade.

5.3.4 Reengenharia

A fase de reengenharia está relacionada à realização de melhorias no processo, no sistema de informação de origem e na qualidade dos dados utilizados para classificar os processos da organização. A reengenharia de um processo envolve o desenvolvimento de um modelo *to-be*, ou seja, o modelo que representa o fluxo de atividades ideal para o processo. Para isso, devem ser utilizados os pontos de melhoria identificados pela etapa de análise e discutidos com as partes interessadas pelo processo.

A identificação de que conjuntos de instâncias fazem parte de cada processo, realizada através do método MANA, permite obter conhecimento a respeito da maneira como os processos da organização deveriam ser tipificados. Dessa forma, iniciativas para aumentar a qualidade de dados da base de origem, com a inclusão de uma tipificação detalhada de processos, são importantes para que os processos da organização deixem de ocorrer de forma desestruturada.

A reavaliação do sistema de informação de origem, que acompanha os processos da organização, também é importante para uma iniciativa de reengenharia que deseje aumentar o grau de estruturação desses processos. Idealmente, um sistema de informação ciente de processo deve ser utilizado, para que seus usuários sejam capazes de identificar a próxima atividade a ser executada a cada passo. Além disso, esse sistema deve ser capaz de controlar o fluxo de atividades de forma não permitir a execução de caminhos indesejados. A adoção de um sistema de BPMS, o desenvolvimento de uma nova ferramenta, ou a reengenharia do sistema existente irá depender do contexto de cada organização.

5.4 Principais Diferenciais

A Tabela 4 resume os principais diferenciais do método MANA em relação às abordagens existentes. Foram comparados somente o framework ProM e o Aris PPM, por serem as ferramentas existentes com um maior número de características em comum com a abordagem deste trabalho: elas suportam, respectivamente, a mineração de processos desestruturados e a análise de processos a partir de uma base padrão contendo atributos de instâncias. Nota-se que nenhum método explícito foi encontrado com as características deste trabalho; na realidade, as ferramentas analisadas suportam implicitamente as etapas do fluxo

proposto pelo método MANA. O sistema desenvolvido com o objetivo de implementar diretamente o fluxo da abordagem proposta será apresentado na seção 5.5.

Tabela 4 – Funcionalidades das ferramentas de mineração de processos

Característica	MANA	ProM	Aris PPM
Suporte à mineração de processos desestruturados	X	X	
Base padrão contendo atributos de processos	X		X
Identificação de processos exploratória e interativa	X		
Clusterização de processos	X	X	
Identificação de instâncias similares	X		
Descoberta de modelos configurável	X	X	
Animação de processos com informações de desempenho	X	X (fuzzy)	
Atividades de reengenharia organizacional	X		
Fluxo de trabalho para analistas não especializados em mineração de processos	X		X

- Suporte à mineração de processos desestruturados

A principal motivação deste trabalho é a mineração de processos desestruturados, ou seja, quando não existe um fluxo bem definido para o processo e existem muitos casos excepcionais. Isso é suportado pelo método MANA em suas etapas de identificação e mineração, com a utilização de uma abordagem de filtragem exploratória de consultas, de técnicas de clusterização, de busca de instâncias similares e de algoritmos de descoberta de modelos com suporte a ruído. O framework ProM, por sua vez, suporta a mineração de processos desestruturados através de técnicas de clusterização e de algoritmos de descoberta de modelos com suporte a ruído. O Aris PPM não prevê a análise de processos desestruturados, exigindo uma forte tipificação dos processos analisados e possuindo somente funcionalidade básica de descoberta de modelos.

- Base padrão contendo atributos de processos

O método MANA propõe a utilização de uma base de dados, contendo todos os dados extraídos do sistema de informação de origem que sejam relevantes para a análise de seus processos. Esta base é o ponto de partida para a seleção de instâncias de processo relacionadas para a mineração. Uma modelagem de dados padrão é utilizada, possuindo os principais dados necessários para auxiliar na identificação de processos. O Aris PPM também armazena dados em uma base padrão, divididos em diversas *dimensões* do processo, porém possui forte tipificação prévia das instâncias.

O framework ProM exige a importação de um arquivo XML contendo as instâncias do processo que será minerado. Porém, para sistemas com processos desestruturados, a extração de um log de eventos específico de um processo não é trivial, sendo frequentemente inviável. Um processo pode estar registrado sob mais de um assunto, e um assunto pode englobar diversos processos. Por exemplo, o processo de contratação de novos empregados de uma organização, caso o sistema tenha suas informações cadastradas em campo de texto livre, pode estar espalhado sob diversos assuntos, tais como: *contratação de empregado*, *contratação de empregada*, *cotratacao de fucionario* (sic), dentre outros. Nesse caso, seria necessário que o usuário gerasse um arquivo XML para cada análise desejada, sendo que não existe informação prévia sobre que processos compartilham um mesmo fluxo.

A exigência de que um usuário comum saiba filtrar informações em um banco de dados também é uma fraqueza da abordagem existente. As ferramentas de extração de logs de eventos para o framework ProM não suportam a filtragem exploratória das instâncias extraídas. Por outro lado, a exportação de uma base completa para um arquivo em formato MXML ou XES, e sua posterior importação no ProM, se torna inviável por motivo de desempenho, da falta de semântica para separar as instâncias carregadas e de filtros que apóiem a abordagem exploratória.

- Identificação de processos exploratória e interativa

A identificação de processos a partir de uma base de dados deveria ser exploratória, permitindo que o usuário utilize o *feedback* de seus resultados para continuamente identificar maneiras de se obter resultados melhores (van der Aalst e Gunther 2007). Para isso, uma abordagem flexível e com a seleção de processos integrada à etapa de mineração se torna necessária. Dessa forma, o método MANA propõe a importação de uma base de instâncias de

processos completa para seu banco de dados. Isso permite que o analista explore as informações existentes e, com seu conhecimento específico sobre o negócio, consiga identificar os processos que serão tratados pelo sistema. Isso é feito a partir da definição de *consultas de processo*. Todos os filtros utilizados para definir uma consulta são armazenados e podem ser modificados pelo usuário a qualquer momento. Os filtros incluem atributos relevantes para a seleção de processos, como assunto, unidades, descrição, data, atividades e interessados. Para cada um desses atributos, é possível ordenar seus valores por frequência. Essa abordagem auxilia o analista de processos a entender em que situações ocorre cada fluxo de processo, o que é importante para adquirir conhecimento a respeito do funcionamento da organização. Embora a exploração dos atributos das instâncias possa ser feita através de consultas SQL na base original, essa situação exige que o usuário tenha conhecimento técnico, não seria facilitada pela interface de uma ferramenta desenvolvida especificamente para isso, exigiria uma posterior extração de dados a cada tentativa e não seria integrada a clusterizações automatizadas.

O framework ProM não suporta a utilização de filtros com atributos que auxiliem na separação de instâncias, como assunto e descrição do processo, durante a análise de um log de eventos, resultando na perda de informações vitais. Pelo contrário, mesmo em algoritmos que lidam com processos desestruturados, assume-se que existe alguma forma de coesão dentro do log. O ProM exige que seja feito um trabalho prévio de seleção de instâncias a partir de uma base de dados. O Aris PPM, embora utilize uma base padrão de processos, supõe que os tipos de processo estejam previamente identificados na base.

- Clusterização de processos

O uso de algoritmos de clusterização é importante quando não é possível separar razoavelmente os fluxos de processos desestruturados somente utilizando filtros manuais. A consideração sobre o que seria um modelo de processo *razoável* deve ser avaliada pelo analista de processos para cada caso. O framework ProM e o método MANA suportam a clusterização de processos. O Aris PPM não considera essa possibilidade.

- Identificação de instâncias similares

O método MANA inclui uma atividade de identificação de instâncias similares. A partir de uma instância escolhida, que seja relacionada a um processo importante para análise,

técnicas de busca de vizinhos próximos de raio fixo permitem agrupar todas as instâncias que possuem características similares a ela. Dessa forma, é possível, em alguns casos, aumentar a precisão da seleção de instâncias utilizadas para a descoberta do modelo de um processo.

- Descoberta de modelos configurável

A descoberta de modelos configurável se refere à utilização de algoritmos de mineração de processos que possam ser adaptados pelo usuário para gerar modelos de processo com diferentes níveis de granularidade e suporte a ruído. Dessa forma, é possível visualizar tanto um modelo do fluxo básico do processo quanto um modelo mais detalhando contendo fluxos excepcionais. Essa característica está presente no framework ProM e no método MANA, que prevê a utilização de técnicas previamente implementadas no ProM.

- Animação de processos com informações de desempenho

O framework ProM possui, somente para o minerador fuzzy, uma animação do fluxo de instâncias ao longo de um modelo de processo. Instâncias de atividades são animadas como pontos ao longo dos fluxos do modelo, a partir da data que a atividade de origem do fluxo começou até a data em que a atividade de destino começou. A animação de processos fornece uma maneira intuitiva de avaliar a evolução de um processo ao longo do tempo e de identificar pontos ineficientes. O método MANA propõe a utilização de uma abordagem similar, com a animação das instâncias filtradas por uma consulta de processo em cima de um modelo gerado a partir da mineração dessas instâncias.

O método propõe ainda a projeção de dados de desempenho sobre as instâncias de uma animação de processos, facilitando a identificação de problemas de desempenho. Dessa forma, é possível visualizar fluxos que são sempre lentos ou, quando for o caso, identificar um caso específico onde uma atividade demorou para ser executada. A instância em questão pode então ser avaliada individualmente, aumentando a possibilidade de entendimento das deficiências do processo. A animação de processos será introduzida em maiores detalhes na seção seguinte, com a apresentação da ferramenta desenvolvida neste trabalho.

Um diferencial da animação suportada pela ferramenta desenvolvida para apoiar o método MANA é a possibilidade de pausar a animação e exibir os detalhes de uma instância específica. A análise de uma instância que desvie do padrão ou tenha sido executada de forma ineficiente é importante para a obtenção de conhecimento respeito de um processo.

Informações como sua descrição, suas partes interessadas e seu fluxo completo permitem o rastreamento da origem de exceções e dos problemas identificados.

- Atividades de reengenharia organizacional

O método MANA inclui em seu fluxo de trabalho três atividades de reengenharia, para que os resultados da análise de processos traduzam em melhorias organizacionais reais. A construção de modelos *to-be* permite que as deficiências identificadas em cada processo sejam corrigidas em novas versões do processo. Essa atividade é suportada pela ferramenta desenvolvida através de um modelador BPMN integrado. Além disso, o método define que a base de dados de origem tenha seus tipos de processo reestruturados, utilizando as consultas de processo como base para tipos pré-definidos, aumentando a qualidade dos dados. Finalmente, a atividade de reengenharia do sistema de origem define que o sistema de informação que suporta os processos seja ciente de processos e seja capaz de utilizar o fluxo desejado de cada tipo de processo para guiar seus usuários.

- Fluxo de trabalho para analistas não especializados em mineração de processos

O framework ProM é a referência indiscutível na área de mineração de processos. Ele exige, porém, um amplo conhecimento de seus algoritmos para que seja utilizado de maneira eficiente (van der Aalst 2011). Usuários que não são especialistas na área se deparam com uma grande quantidade de técnicas e configurações que inibem o uso da ferramenta. Dessa forma, um dos objetivos do método MANA é fornecer, ao mesmo tempo, o estado da arte em mineração de processos e facilitar seu uso por analistas do negócio que, mesmo não sendo especialistas no assunto, buscam aprimorar os procedimentos internos de suas organizações. Seu objetivo não é substituir o framework ProM, mas se inspirar nele para facilitar o acesso de organizações ineficientes e pouco estruturadas a técnicas de mineração de processos.

5.5 Ferramenta Desenvolvida

Para suportar o método MANA, foi desenvolvida uma ferramenta de mesmo nome que suporta a maioria das atividades presentes no fluxo de trabalho proposto. A Figura 41 mostra a interface da ferramenta. O restante desta seção está dividido da seguinte forma. A seção 5.5.1 apresenta os requisitos funcionais da ferramenta. A seção 5.5.2 introduz as principais tecnologias utilizadas. A seção 5.5.3 ressalta o rastreamento entre os módulos desenvolvidos e as atividades do método proposto. As seções 5.5.4 a 5.5.12 detalham cada um desses

módulos. Os modelos de processo apresentados durante a descrição da ferramenta são apenas ilustrativos, não correspondendo necessariamente a resultados provenientes de análises reais.

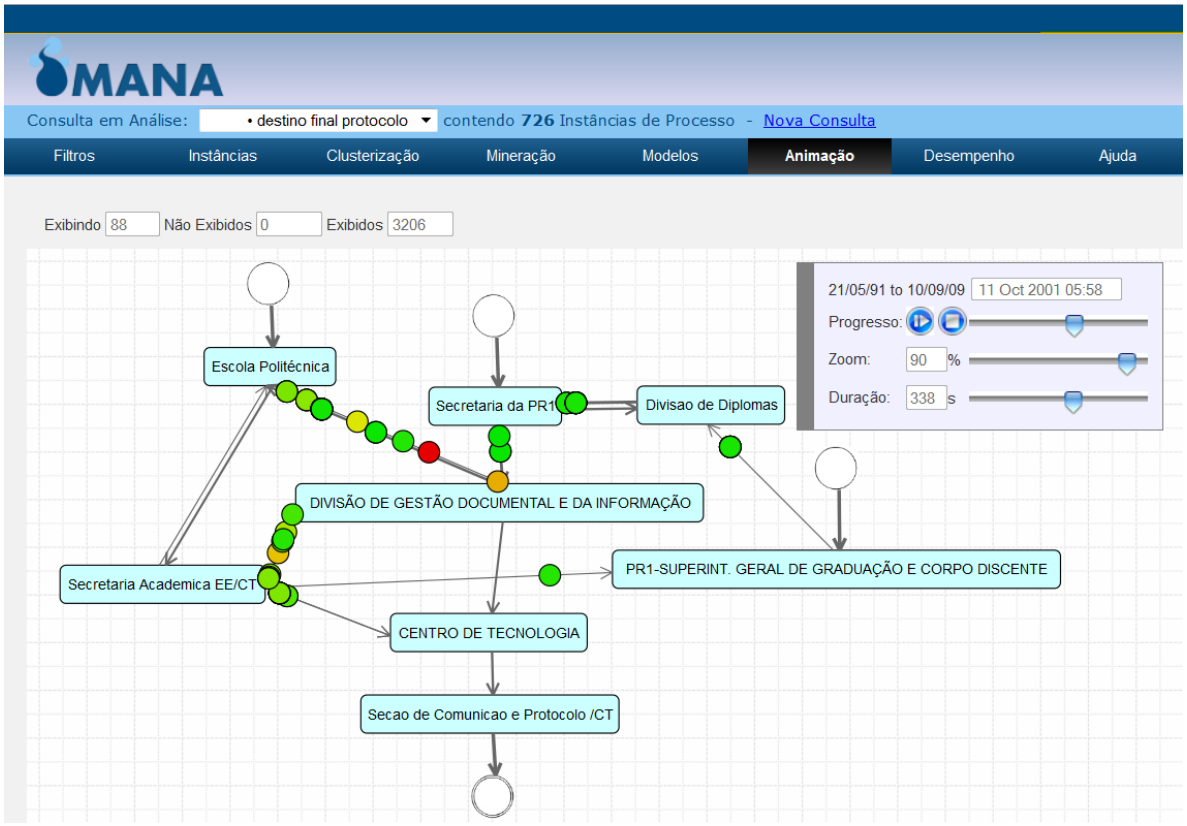

Figura 41 – Ferramenta desenvolvida para suportar o método MANA

5.5.1 Requisitos Funcionais

Os requisitos funcionais da ferramenta, tendo como base os conceitos apresentados nas seções anteriores e o fluxo de trabalho proposto pelo método MANA, são:

- RF1 - O sistema deve possuir uma base de dados padrão de instâncias de processo.
- RF2 - O sistema deve permitir o gerenciamento de consultas de processo.
- RF3 - O sistema deve permitir a existência de uma hierarquia de consultas de processo.
- RF4 - O sistema deve permitir a exportação das instâncias contidas em uma consulta de processo para a utilização com ferramentas externas.
- RF5 - O sistema deve permitir a exportação de modelos de processo para a utilização com ferramentas externas.

- RF6 - O sistema deve permitir a exploração da base de dados padrão, através de buscas textuais sobre seus atributos.
- RF7 - O sistema deve permitir a criação de filtros sobre as instâncias de uma consulta de processo, a partir dos valores resultantes de uma busca sobre os atributos da base.
- RF8 - O sistema deve permitir a clusterização das instâncias contidas em uma consulta de processo.
- RF9 - O sistema deve permitir a seleção de instâncias similares a uma instância selecionada.
- RF10 - O sistema deve permitir a visualização das instâncias individuais contidas em uma consulta de processo.
- RF11 - O sistema deve permitir a descoberta de modelos de processo a partir de uma consulta, utilizando algoritmos de mineração de processos.
- RF12 - O sistema deve permitir a visualização e edição de modelos de processo.
- RF13 - O sistema deve permitir a visualização de uma animação de processos a partir de um modelo de processo e de sua consulta correspondente.
- RF14 - O sistema deve permitir a visualização de informações de desempenho de cada instância de atividade exibida em uma animação de processos.
- RF15 - O sistema deve permitir, para uma consulta de processo, a visualização e a exportação de relatórios de desempenho para as atividades e unidades pertinentes à consulta, contendo informações agregadas a respeito do tempo de execução de cada atividade.
- RF16 - O sistema deve permitir, para uma consulta de processo, a visualização de gráficos que permitam ao usuário visualizar informações de atraso relacionadas a cada instância de atividade ou cada passagem por uma unidade.

5.5.2 Tecnologias Utilizadas

A ferramenta foi desenvolvida em Java como um sistema Web, tendo como base o framework JSF (JavaServer Faces) 2.0 (Oracle 2009). A base de dados, derivada da modelagem apresentada na seção 5.2, foi implementada utilizando o sistema de

gerenciamento de banco de dados SQL Server 2008 R2 (Microsoft 2010). As seguintes bibliotecas e ferramentas externas suportam diretamente as atividades do método:

- Weka (Hall et al. 2009), biblioteca de mineração de dados, utilizada para a clusterização de consultas com o algoritmo k-means;
- Raphaël (Baranovskiy 2012), biblioteca em JavaScript com suporte ao desenho de gráficos SVG (*Scalable Vector Graphics*), para a construção dos módulos de modelagem e de animação de processos;
- OpenXES (Günther 2009), implementação do padrão XES, usado como interface para a transferência de instâncias de processo aos algoritmos de mineração de processos.
- As implementações do minerador de heurísticas, utilizado para a descoberta de modelos, e do *Disjunctive Workflow Schema*, utilizado para a clusterização de consultas, foram adaptadas do framework ProM (Van Dongen et al. 2005);
- GraphViz (Graphviz 2011), utilizado para o layout dos modelos de processo gerados, sendo que sua interface para códigos em Java foi encontrada em Szathmary (2011).

5.5.3 Suporte dos módulos desenvolvidos às atividades do método MANA

A ferramenta desenvolvida suporta a maioria das atividades propostas pelo método MANA. As atividades de *reestruturar tipos de processo na base de origem* e *reavaliar sistema de informação original* estão fora do escopo da ferramenta, devendo ser realizadas externamente. Os seguintes módulos foram implementados: *cadastro de consultas, filtros, clusterização, instâncias, mineração, modelagem, animação e análise de desempenho*. Eles serão apresentados nas seções seguintes deste trabalho. A Tabela 5 mostra como cada atividade do método é rastreada para cada módulo desenvolvido.

Tabela 5 – Rastreamento entre as atividades do método MANA e os módulos da ferramenta desenvolvida

Atividade	Módulo
Iniciar consulta de processo	Cadastro de consultas de processo
Filtragem exploratória	Filtros
Clusterizar consulta	Clusterização

Incluir instâncias similares	Instâncias
Minerar consulta	Mineração
Validar modelo de processo *as-is*	Modelagem
Analisar animação de instâncias	Animação
Analisar relatórios de desempenho	Análise de desempenho
Analisar instâncias individualmente	Instâncias Animação Análise de desempenho
Realizar reengenharia em modelo de processo *to-be*	Modelagem

5.5.4 Cabeçalho

O cabeçalho inclui o logo do sistema, que dá acesso à página principal. A *consulta* atualmente em análise é exibida. Ela é uma pasta de trabalho, representando o projeto atualmente em desenvolvimento que agrupa um conjunto de instâncias relacionadas. A seleção de uma consulta ativa todos módulos do sistema, que passam a trabalhar com seus dados e ficam disponíveis através do menu principal. O número de instâncias contidas atualmente na consulta é exibido no cabeçalho. O usuário pode alternar entre consultas a qualquer momento, utilizando a caixa de seleção na parte superior do sistema. No mesmo local, é possível selecionar um *workspace* para trabalhar, sendo que cada *workspace* representa uma base de dados de origem importada para a base padrão de processos.

Figura 42 – Cabeçalho do sistema

5.5.5 Cadastro de Consultas de Processo

Este módulo é a tela inicial do sistema, e representa o início do fluxo de trabalho. Ele apresenta uma visão geral das consultas de processo, permitindo seu cadastramento, alteração e exclusão. Quando uma consulta é selecionada, os módulos do sistema são ativados no menu principal. Esses módulos passam a utilizá-la como sua pasta de trabalho, ou seja, trabalhar em cima das instâncias filtradas pela consulta. O sistema permite a exportação das instâncias filtradas por uma consulta de processo para o formato MXML. Isso é importante para usuários avançados, caso desejem aprofundar sua análise utilizando técnicas que ainda não se encontram implementadas na ferramenta. O formato MXML foi escolhido ao invés do padrão XES porque, embora seja mais antigo, ele pode ser importado pelas duas versões atuais do framework ProM (5.2 e 6.1). Nota-se que existe uma hierarquia entre consultas, sendo que as consultas filhas utilizam as instâncias de seus pais como ponto de partida para as atividades de filtragem.

Figura 43 – Cadastro de consultas de processo

5.5.6 Filtros

O módulo de filtros permite a exploração da base de dados e a associação de filtros a uma consulta. Ele é dividido em três partes: busca de novos filtros, filtros atuais e filtros da consulta pai. A busca por novos filtros tem por objetivo analisar os valores presentes para cada atributo das instâncias de processo, e selecionar os valores desejados para a consulta atual. A busca por atributos pode incluir um valor, ou ser feita uma busca vazia que retorna todos os valores do atributo para a consulta atual. Por exemplo, a Figura 44 ilustra uma busca por *descrições* contendo o texto *diploma* em uma consulta. Os valores retornados são ordenados pelo número de instâncias, métrica que auxilia na identificação de valores de alta importância. Uma busca vazia por *assuntos*, por exemplo, fornece um ponto de partida para a análise dos processos mais importantes da base de dados. O símbolo % pode ser usado no

meio do texto de uma busca para indicar qualquer número de caracteres, como, por exemplo, *dipl%gradua*.

Figura 44 – Exploração de atributos e seleção de filtros

Os resultados de uma pesquisa de atributos podem ser selecionados e incluídos como filtros da consulta. Dessa forma, o conjunto de instâncias da consulta passa a se restringir aos valores selecionados. Os filtros podem também ser salvos como uma consulta aninhada, formando uma hierarquia de consultas cada vez mais específicas. Os filtros atuais são exibidos em uma listagem. Eles são de extrema importância, pois definem as instâncias de processo que serão utilizadas pelos demais módulos do sistema. Os filtros da consulta pai, que esteja um nível acima na hierarquia de consultas, também são exibidos em uma segunda listagem. Filtros incluem o *atributo* que está sendo filtrado; uma comparação, que pode ser *igual a ou diferente de;* e o valor do atributo que está sendo considerado no filtro. Filtros diferindo somente no valor são unidos logicamente por cláusulas OR. Cada um destes agrupamentos é unido por cláusulas AND, tornando a consulta uma expressão na forma normal conjuntiva.

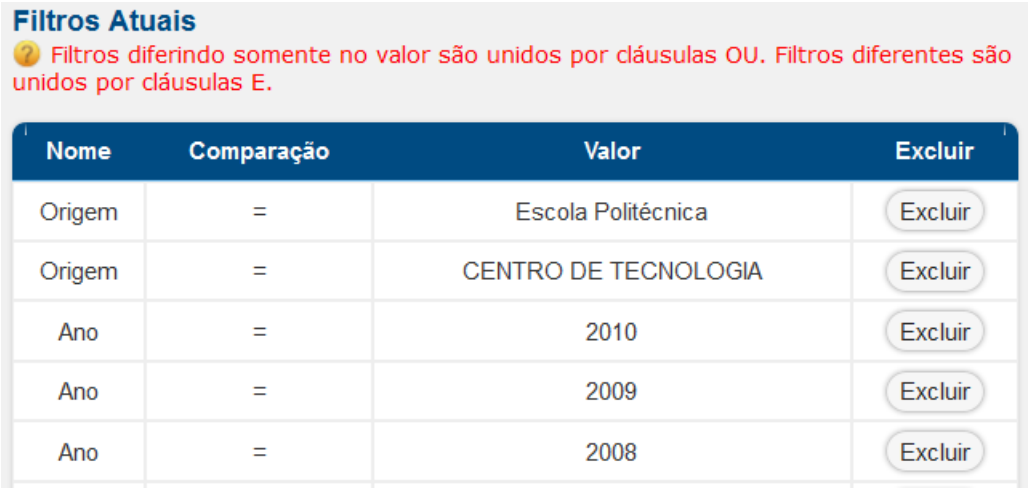

Figura 45 – Filtros atuais

5.5.7 Clusterização

O módulo de clusterização deve ser utilizado quando não é possível obter um modelo de processo considerado razoável pelo analista utilizando somente a filtragem manual. São utilizadas duas técnicas: a clusterização DWS e a clusterização de unidades, que constrói um perfil de unidades similar à abordagem utilizada pelo algoritmo de clusterização de *traces*. Ambos utilizam o algoritmo k-means para a clusterização, exigindo como entrada o número desejado de clusters. Os algoritmos de clusterização foram discutidos em maiores detalhes na seção 3.3 deste trabalho.

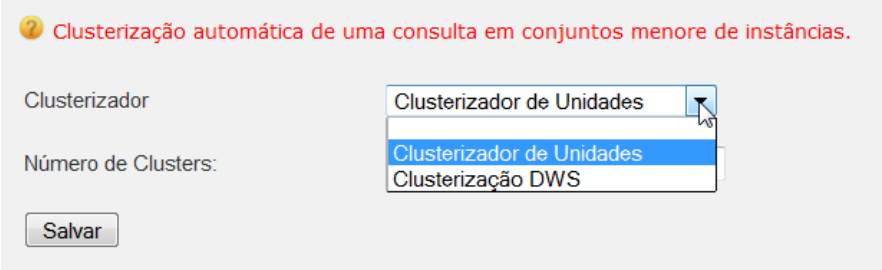

Figura 46 – Módulo de clusterização

5.5.8 Instâncias

O módulo de instâncias permite a visualização de detalhes relacionados às instâncias contidas na consulta selecionada, como seus atributos e seu fluxo. Este módulo permite ainda o terceiro tipo de seleção de instâncias de processo do método MANA, a inclusão instâncias similares. Para isso, o usuário deve primeiro identificar uma instância relacionada ao processo que se deseja minerar. Dessa forma, o sistema possibilita a inclusão de instâncias similares à instância selecionada, utilizando técnicas de busca de vizinhos próximos de raio fixo (tradução livre de *fixed-radius near neighbors*). Atualmente, somente um método se encontra implementado: o *fingerprint* da descrição das instâncias.

O método de *fingerprint* utiliza uma abordagem de identificação de registros duplicados, baseada no método descrito pela ferramenta Google Refine (Morris 2012). Para a descrição de cada instância, são realizados os passos abaixo. O *fingerprint* resultante é armazenado na base de dados. O resultado é uma *string* normalizada que elimina uma grande quantidade de variações no cadastramento textual de dados. É possível configurar a distância desejada para o *fingerprint* da instância selecionada, utilizando a distância de edição entre as duas *strings.* Maiores informações sobre técnicas de identificação de registros duplicados podem ser encontradas em Elmagarmid et al. (2007).

- Alteração de todas as letras para maiúsculo.
- Remoção de acentuação.
- Remoção de espaços no início ou final da *string*.
- Remoção de caracteres especiais e de espaços duplicados.
- Remoção de palavras comuns ou *stopwords* (de, os, com, e, em, um, etc).
- Aplicação do Stemmer de Porter (extração do radical de cada palavra).
- Remoção de radicais duplicados.
- Ordenação das palavras.

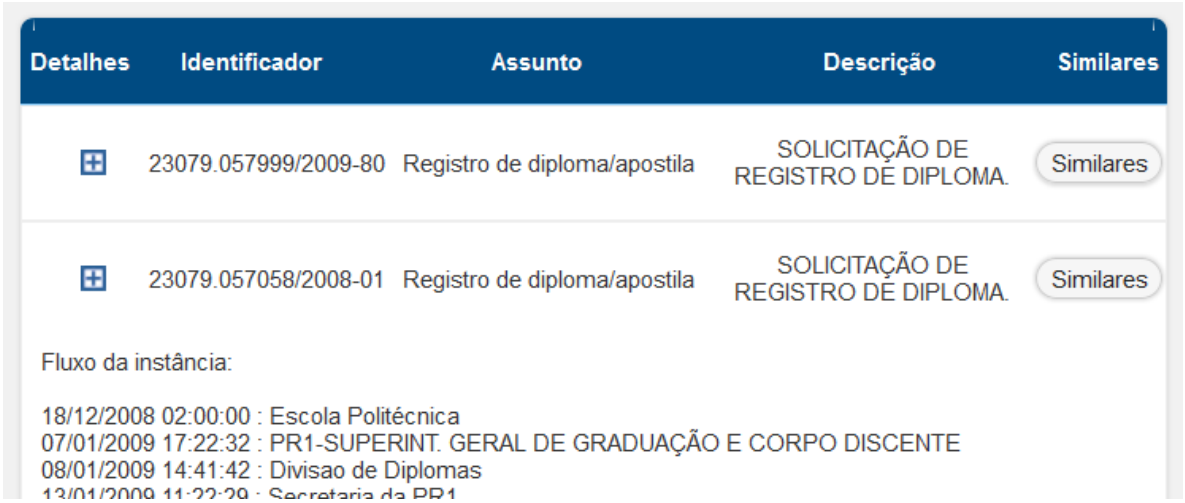

Figura 47 – Módulo de visualização de instâncias

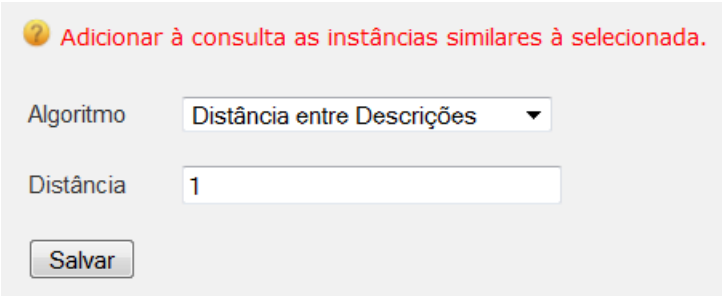

Figura 48 – Inclusão de instâncias similares na consulta

5.5.9 Mineração

O módulo de mineração tem como objetivo descobrir um modelo de processo a partir das instâncias selecionadas em uma consulta de processo. Atualmente, somente o minerador de heurísticas é suportado. Ele foi escolhido por ser robusto à existência de ruído, muito comum em sistemas onde a entrada de dados é feita manualmente. Como muitos sistemas de protocolo, como os estudados neste trabalho, armazenam por onde os processos passaram, e não as atividades que foram executadas, este módulo permite gerar tanto o fluxo de atividades do processo quanto o fluxo entre unidades organizacionais.

O sistema permite a configuração pelo usuário de dois parâmetros do algoritmo: o limiar de dependência e a utilização ou não da heurística de todas as atividades conectadas. Estes parâmetros foram considerados como sendo de fácil entendimento pelo usuário. A variação do limiar de dependência é importante durante a atividade de exploração do processo, permitindo alternar entre a modelagem somente dos comportamentos mais

frequentes e de todos os comportamentos executados no mundo real. O segundo parâmetro permite que o usuário opte por visualizar ou não as atividades/unidades que são pouco frequentes no processo considerado. Para os demais parâmetros, valores padrão são utilizados de maneira transparente.

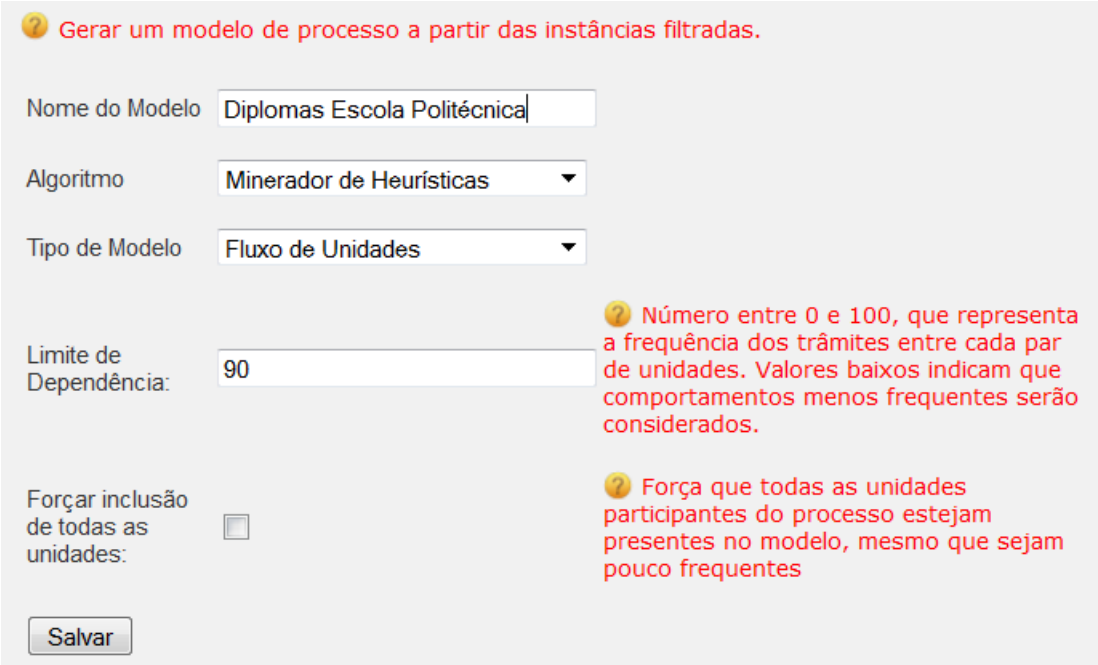

Figura 49 – Módulo de mineração

Um problema encontrado em sistemas de protocolo é a ênfase dada no registro das unidades organizacionais por onde um processo tramita, e não nas atividades executadas por estas unidades. A descrição das tarefas geralmente possui semântica fraca ou mesmo ausente. Dessa forma, o sistema desenvolvido permite a mineração de modelos para o fluxo entre unidades organizacionais, além do fluxo de atividades do processo. O módulo de modelagem permite que posteriormente o usuário inclua as atividades executadas por cada unidade do modelo, realizando as substituições necessárias. Apesar de a utilização das unidades executoras ao invés das atividades executadas na construção de modelos de fluxo não seja ideal (uma unidade pode executar mais de uma atividade, e uma atividade pode ser executada por mais de uma unidade), ela permite o aproveitamento de técnicas de mineração de processos em casos onde seu uso seria impossibilitado pela ausência de atividades bem definidas.

5.5.10 Modelagem

O sistema inclui um modelador de processos simplificado que implementa os principais elementos da linguagem BPMN. Este modelador tem como objetivo suportar a visualização dos modelos gerados pelo módulo de mineração, sua validação pelas partes interessadas e a construção de modelos *to-be*. Caso necessário, um modelo pode ser refinado, com a inclusão, edição e ordenação de seus elementos.

A Figura 50 ilustra o módulo de modelagem de processos. Cada modelo está ligado à consulta de processo que deu origem a ele, para fins de organização de dados. É possível alterar o nível de zoom da visualização e organizar automaticamente os elementos do modelo, utilizando o algoritmo descrito em (Gansner et al. 1993) e implementado pela ferramenta GraphViz. A espessura das setas exibidas indica o número de vezes que esta relação apareceu nas instâncias que geraram o modelo de processo. Dessa forma, uma seta muito forte representa um fluxo muito frequente no processo.

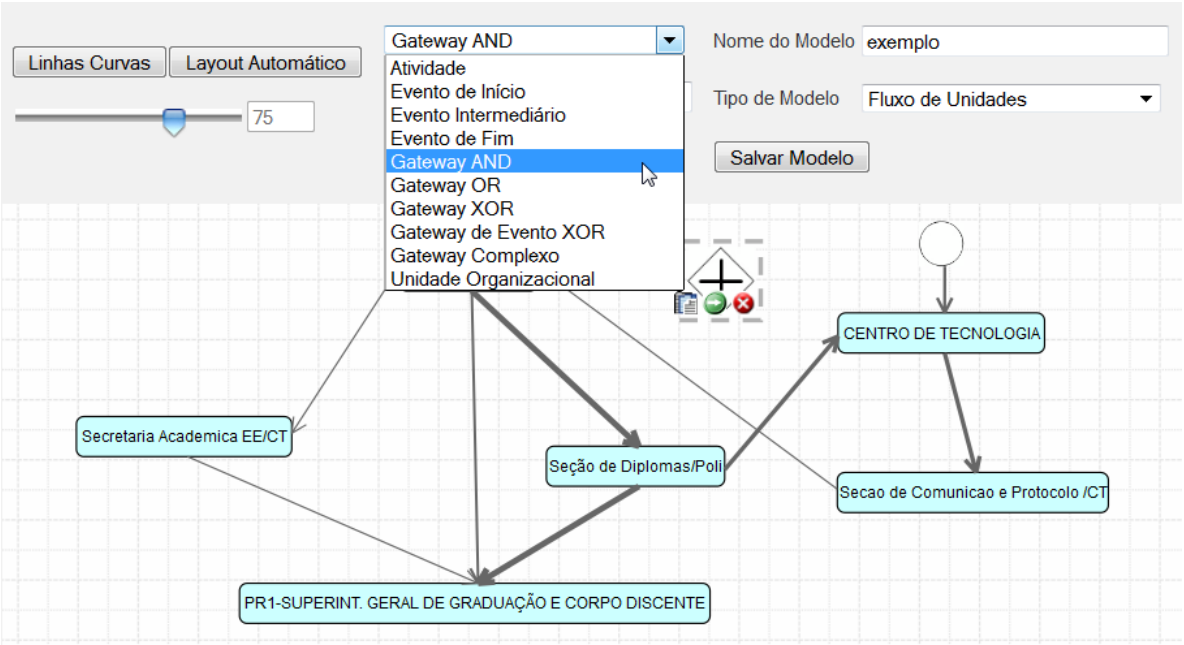

Figura 50– Módulo de modelagem

Os elementos suportados são: atividades, fluxos de sequência, eventos (de início, intermediários e de fim) e *gateways* (AND, OR, XOR, evento XOR e complexo). Um elemento não existente na linguagem foi adicionado: unidade organizacional. Seu objetivo é suportar a modelagem de fluxos entre unidades, necessária quando a base de dados utilizada não inclui informações sobre as atividades executadas por cada unidade.

Uma unidade organizacional é representada da mesma forma que uma atividade, mas com a cor azul. Cada unidade pode, através de entrevistas com as partes interessadas do processo, ser substituída no modelo pela atividade (ou atividades) que ela executa. Dessa forma, um modelo de processo tradicional é obtido. Embora a notação BPMN utilize *swimlanes* para separar os diferentes executores de um processo, esta estrutura não é capaz de se adequar aos requisitos necessários para modelar um processo minerado a partir de um fluxo de unidades organizacionais.

A inclusão de um módulo de edição de processos é importante para que os refinamentos necessários possam ser feitos de maneira integrada nos modelos resultantes da etapa de mineração. Sua utilização também é vital durante a reengenharia, permitindo a construção da versão *to-be* de cada processo. O desenvolvimento de um modelador de processos robusto, porém, está fora do escopo deste trabalho. Caso seja necessário, os modelos gerados podem ser exportados no formato XPDL (*XML Process Definition Language)* para utilização em ferramentas comerciais de modelagem de processos. A exportação de modelos também é importante quando a organização já utiliza tradicionalmente outra ferramenta de modelagem.

5.5.11 Animação

O módulo de animação foi construído tendo como inspiração a animação de processos implementada pelo minerador fuzzy no framework ProM e ferramentas de simulação de processos. Como todos os eventos registrados para uma instância possuem uma *timestamp* associada, eles são ordenados e projetados sobre o modelo de processo, resultando em um vídeo do processo. Esta abordagem permite visualizar de maneira convincente os problemas reais do processo *as-is*, ao contrário da simulação, cujos resultados podem ser contestados (van der Aalst 2011).

A ferramenta desenvolvida expande a proposta do framework ProM de duas formas: interatividade e sobreposição de informações de desempenho de atividades. Enquanto a abordagem existente exibe os nós de forma anônima, o módulo de animação permite pausá-la e avaliar os detalhes de cada instância específica, exibindo todo seu fluxo, sua posição atual no fluxo e demais informações cadastradas no sistema.

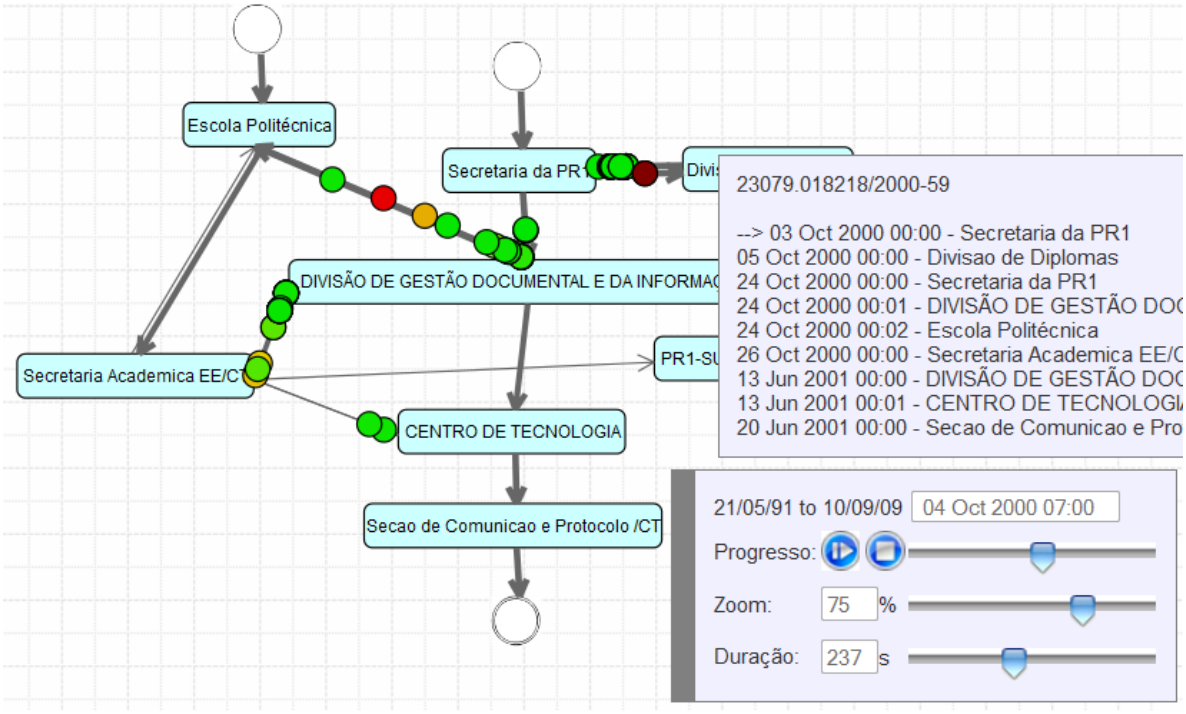

Figura 51 – Módulo de animação

O módulo calcula a duração de cada evento (instância de atividade) a partir do tempo decorrido entre o início do evento e o início do evento posterior. A partir destes valores, uma escala de cores entre vermelho, amarelo e verde é obtida no espaço de cores HSV (matiz, saturação e valor), sendo que o vermelho (0, 1, 0.9) é atribuído ao evento mais lento, e o verde (120, 1, 0.9) é atribuído ao evento mais rápido. Esse código de cores fornece uma informação extra que permite identificar intuitivamente as atividades com baixo desempenho e todos os casos em que houve lentidão.

A representação de informações de desempenho em um código de cores foi inspirada em outro plug-in do framework ProM. Ele agrega informações de desempenho de cada atividade do modelo entre três cores, verde, amarelo e vermelho. Porém, não atribui essas cores a instâncias individuais, e não exibe a evolução do processo ao longo do tempo.

5.5.12 Análise de Desempenho

A análise de desempenho tem como objetivo gerar relatórios de desempenho de cada unidade organizacional que participa do processo consultado. Uma abordagem similar poderia ser realizada para o desempenho das atividades do processo, porém não se encontra implementada atualmente. Para cada item do relatório, exibe-se o número de instâncias das

quais a unidade participou, o número de eventos analisados e não analisados (não é possível analisar o atraso de um evento que não possua um evento seguinte), e os atrasos mínimo, médio e máximo desta unidade para as instâncias consideradas. O relatório pode ser exportado em uma planilha para armazenamento e atividades posteriores. O objetivo atual desse módulo é fornecer uma visão direta e intuitiva do desempenho do processo. O suporte futuro a um armazém de dados de processos, como proposto por Casati et al. (2007), possui o potencial de fornecer uma análise mais aprofundada, mas se encontra fora do escopo deste trabalho.

Figura 52 – Análise de desempenho

Para auxiliar na análise de desempenho, dois tipos de gráfico são suportados nesse módulo. O gráfico de dispersão de atrasos tem como objetivo plotar o tempo que a unidade levou para executar cada instância de atividade. A medida calcula a diferença entre a data de saída e a data de chegada na unidade. Isso permite avaliar detalhadamente os dados consolidados pela tabela acima. Cada ponto do gráfico possui um *hyperlink* para a instância relacionada, permitindo identificar quais foram os casos ineficientes do processo. Na Figura 53, o eixo x indica a data, e o eixo y o atraso da instância na unidade, em dias.

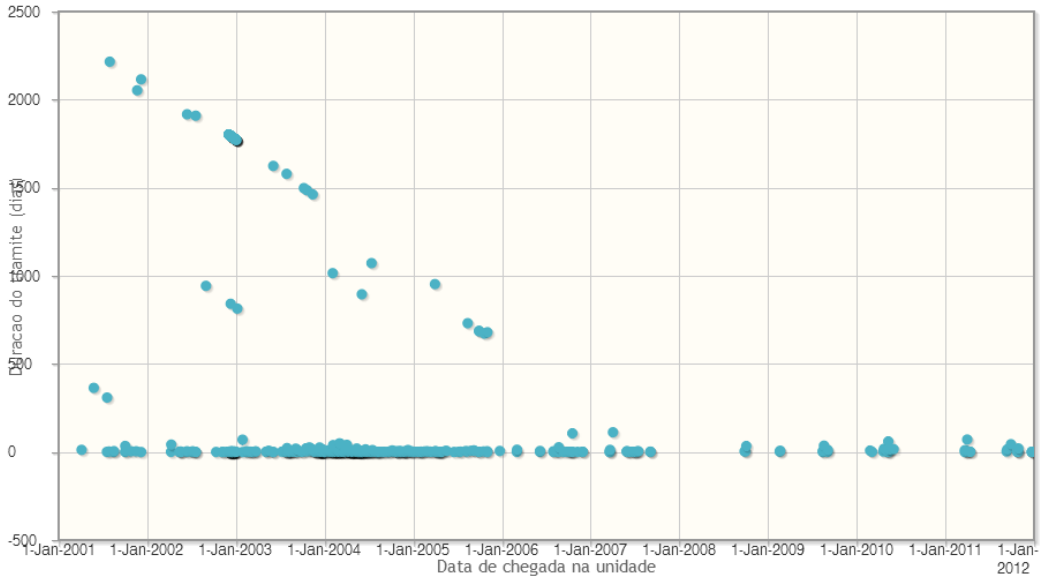

Figura 53 – Gráfico de dispersão de atrasos

O sistema disponibiliza ainda um gráfico de linha do tempo, ligando os pontos relacionados às datas de chegada e de saída de cada instância de processo de uma unidade selecionada. O eixo y deste gráfico representa apenas uma indicação sequencial dos casos plotados. Esta visualização permite avaliar a vazão de uma unidade, com a identificação dos momentos em que ela possui muitas instâncias acumuladas. Uma abordagem similar é proposta por van der Aalst (2011). Vale ressaltar ainda que é comum a existência de momentos em que diversos casos são tratados de uma vez; por exemplo, na Figura 54, uma grande quantidade de processos foi liberada da unidade, ao mesmo tempo, em julho de 2010. Essa situação pode exigir análises posteriores.

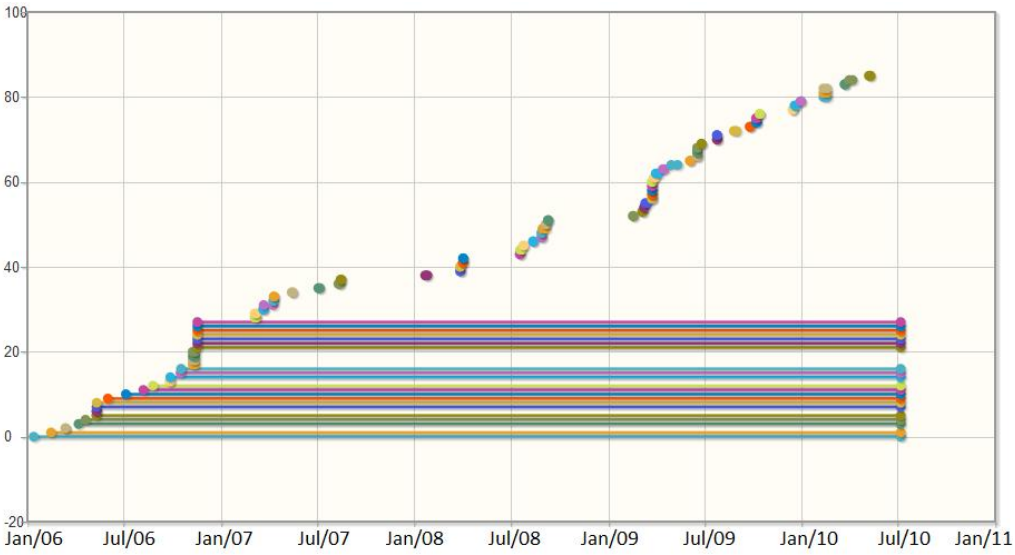

Figura 54 – Gráfico de linha do tempo para uma unidade

5.6 Considerações finais

Este capítulo apresentou os conceitos utilizados pelo método MANA e descreveu o fluxo de trabalho proposto por este trabalho. A ferramenta desenvolvida para suportar as atividades propostas também foi detalhada. O capítulo seguinte tem como objetivo apresentar duas situações da aplicação do método proposto. Em especial, serão utilizadas bases de dados da Universidade Federal do Rio de Janeiro e do Ministério do Planejamento, Orçamento e Gestão. Será mostrado como a aplicação do fluxo de trabalho apresentado, suportado pela ferramenta desenvolvida, supera os resultados obtidos através das abordagens existentes.

Capítulo 6 – Provas de conceito

Esta seção apresenta duas provas de conceito de mineração e análise de processos de negócio utilizando o método MANA. Seu objetivo é exemplificar a utilização da abordagem desenvolvida e ressaltar como a utilização de ferramentas alternativas dificultaria a análise, tornando-a mais imprecisa e ineficiente. A seção 6.1 apresenta uma prova de conceito a partir de dados extraídos do Sistema de Controle de Processos e Documentos do Ministério do Planejamento, Orçamento e Gestão do Brasil, que gerencia os processos pertinentes ao Ministério. Na seção 6.2, outra prova de conceito analisa o Sistema de Acompanhamento de Processos da Universidade Federal do Rio de Janeiro. As duas bases de dados analisadas possuem em comum o fato de seus processos serem desestruturados, sem a formalização do fluxo de trabalho que deve ser seguido, com a possibilidade de entrada de dados de forma textual em campo livre e com métodos de classificação de processos muito genéricos.

Ambas as bases de dados analisadas são originárias de sistemas de controle de processos de organizações públicas brasileiras. Este tipo de sistema foi uma grande motivação para este trabalho, pois registram grande parte dos processos de uma organização, mas, na maioria dos casos, não controlam o fluxo de atividades nem a entrada de dados pelos usuários. Isso faz com que a utilização de técnicas de mineração de processos desestruturados seja adequada para lidar com os dados desses sistemas. E, além de objetivar a obtenção de modelos de processo corretos, esta análise pretende causar impacto para a desestruturação dos processos das organizações, motivando projetos de reengenharia de processos.

Como o ProM foi a única abordagem encontrada com algum suporte a processos desestruturados, ele foi selecionado como ponto de comparação para as provas de conceito abaixo. O Aris PPM, por exemplo, exige uma forte tipificação dos processos e não suporta a mineração de processos com ruído, não sendo capaz de lidar com as situações estudadas.

6.1 Controle de Processos e Documentos do Ministério do Planejamento

O Sistema de Controle de Processos e Documentos – CPROD Web é responsável por gerenciar os processos e documentos pertinentes ao Ministério do Planejamento, Orçamento e Gestão do Brasil. Ele inclui atividades de recebimento, registro, cadastramento, tramitação, expedição, classificação, destinação e autuação desses processos (Ministério do Planejamento, Orçamento e Gestão 2004). Esta prova de conceito tem como objetivo analisar a base de dados do CPROD do ponto de vista da mineração de processos. No momento da análise, o sistema possuía mais de 3 milhões de processos cadastrados e 16 milhões de trâmites (eventos), o que mostra a grande dimensão da base de dados utilizada. O escopo deste trabalho se limita à análise de somente um processo, porém a mesma abordagem pode ser utilizada para avaliar o restante dos processos da base.

6.1.1 Estrutura da Base de Dados

Nesta seção serão apresentados os principais conceitos do CPROD e os relacionamentos entre eles, com base nas informações contidas no manual do sistema (Ministério do Planejamento, Orçamento e Gestão 2004). Seu objetivo não é detalhar a modelagem de dados completa do sistema, que possui uma grande quantidade de tabelas e atributos, mas sim descrever as informações utilizadas neste trabalho. Serão analisados somente os processos que tiveram a participação da Secretaria de Logística e Tecnologia da Informação. A figura abaixo apresenta um modelo de dados lógico simplificado do CPROD. O código identificador de cada entidade é gerado automaticamente pelo sistema; os demais atributos são relacionados ao negócio.

A entidade central do CPROD é o *protocolo*. Ela registra processos e documentos de interesse do Ministério. Um protocolo possui uma *data de abertura*, um *número* e um *assunto*. O assunto é um campo de texto livre, que atua como a descrição do protocolo. A entidade *classe de temporalidade* implementa um código de classificação por assuntos. Uma classe de temporalidade está ligada a um protocolo através da tabela *temporalidade*. Um protocolo possui ainda uma *procedência*, que é a unidade que originou o protocolo, através de um relacionamento direto entre as duas entidades. A tabela intermediária de *procedência do protocolo* registra as partes interessadas de um protocolo e seus solicitantes.

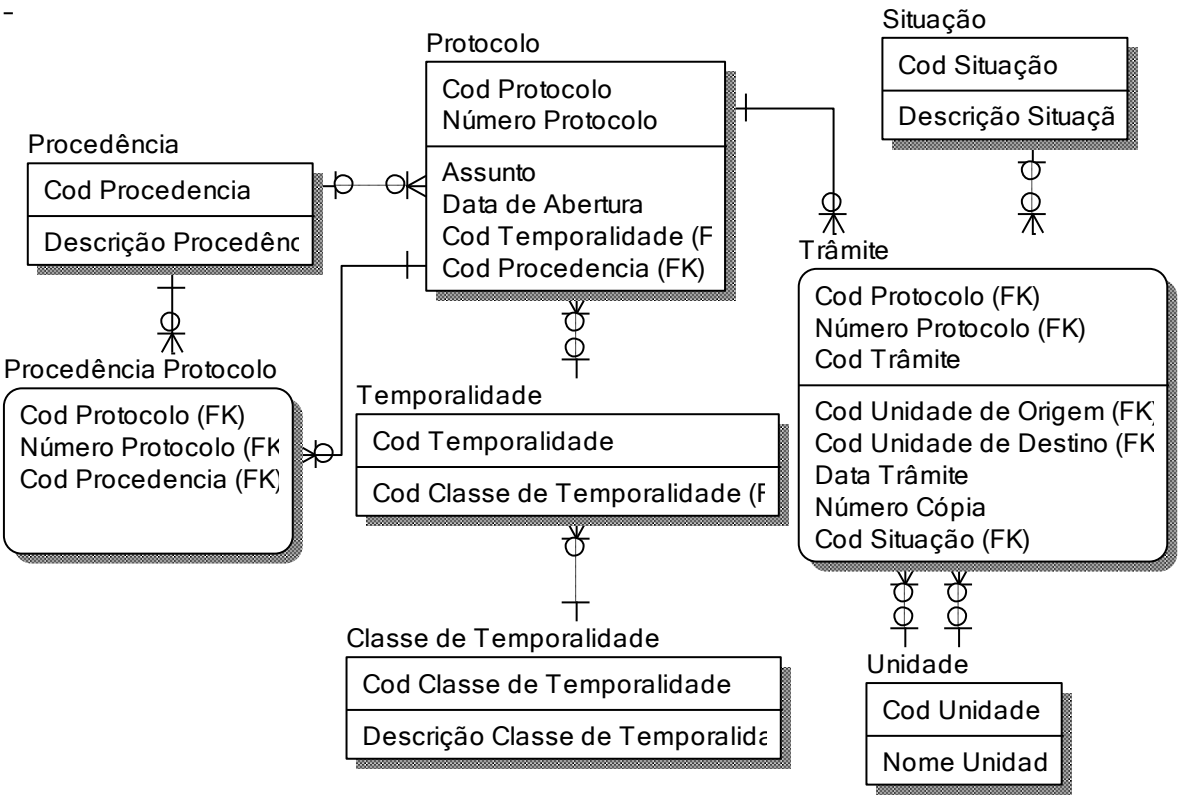

Figura 55 – Modelo de dados lógico simplificado do sistema CPROD Web

A entidade *trâmite* registra o andamento de um protocolo de uma unidade de origem para uma unidade de destino. Um trâmite pode ser de um protocolo original ou de uma cópia; os trâmites de cópia não foram carregados para a base da ferramenta MANA, já que não fazem parte do fluxo principal do processo. Um trâmite registra ainda a data em que o protocolo foi enviado. Embora o trâmite seja a entidade mais adequada para a identificação das *instâncias de atividade* de um processo, o sistema não registra exatamente o que é feito em cada etapa, somente seus atores. Dessa forma, a análise realizada neste trabalho possui enfoque na identificação do fluxo dos processos entre unidades organizacionais, e não das atividades executadas por elas. Nota-se ainda que o CPROD não está ciente do fluxo de trâmites entre unidades para cada protocolo, permitindo que o andamento de processos similares ocorra de maneiras diferentes, dependendo de seus executores. O resultado disso são processos altamente desestruturados. Essa abordagem também dá margem a erros de cadastramento.

A tabela abaixo mostra como os conceitos da base de dados do CPROD foram mapeados para a ferramenta MANA. Esse mapeamento deve ser realizado através do estudo da correspondência entre os conceitos da base de dados de origem e da base da ferramenta MANA. A operação de migração de dados foi feita através de uma carga ETL. Nota-se que o atributo de assunto de um protocolo é mapeado para a descrição de uma instância de processo na ferramenta MANA. Isso foi feito porque a classe de temporalidade possui maior semelhança com um assunto estruturado, enquanto que o atributo de assunto, digitado manualmente em campo de texto livre, se assemelha a uma descrição do protocolo.

Tabela 6 – Mapeamento entre conceitos do CPROD Web e do MANA

CPROD	MANA
Protocolo	Instância de processo
Número do Protocolo	Identificador da Instância
Data de abertura do protocolo	Data de início da instância
Assunto do protocolo	Descrição da instância
Classe de temporalidade	Assunto
Trâmite	Evento
Unidade	Unidade
Status do trâmite mais recente do protocolo	Status da instância
Procedência	Origem
Procedências do protocolo	Interessados

Embora o CPROD suporte uma classificação estruturada por assuntos, através das tabelas de *temporalidade*, essa informação está vazia para aproximadamente 2/3 dos protocolos presentes na base. Dessa forma, o atributo *assunto* de um protocolo, embora pouco estruturado, possui maior utilidade na prática do que a temporalidade. O campo de assunto também possui, geralmente, informações mais detalhadas, permitindo uma seleção de instâncias de processo mais precisa.

Como esse atributo é cadastrado em campo de texto livre no sistema, a busca de processos por assunto encontra diversos desafios. A liberdade permitida ao usuário faz com que processos de um mesmo tipo possuam assuntos muito diferentes. Distinções comuns incluem a utilização de sinônimos, abreviaturas, inversão de palavras, utilização ou não de acentuação e pontuação, erros de digitação, dentre outras. Foram encontradas ainda duas abordagens para o cadastramento de assuntos: a utilização de um tipo de processo genérico (e.g. *relatório de disponibilidade de rede*), e o detalhamento com informações específicas do processo (e.g. *relatório de disponibilidade de rede, de 2010 a 2011, segundo o ofício XXX*). Foi realizada uma tentativa de aumentar a qualidade de dados de cadastramento de assuntos utilizando o método de *fingerprint* discutido anteriormente. Porém, como não foi possível validar as alterações com as partes interessadas pelos processos da base, elas ainda não foram aplicadas.

6.1.2 Mineração e Análise dos Processos

A mineração de um processo utilizando a ferramenta MANA é iniciada com a criação de uma consulta. Uma consulta inicia contendo todas as instâncias da base do sistema, que são reduzidas sucessivamente com a utilização de filtros. Devido à abrangência da base de dados, o escopo da análise foi reduzido para conter somente os processos que passaram pelas unidades subordinadas à Secretaria de Logística e Tecnologia da Informação – SLTI. Isso foi feito através de uma busca pelo atributo *unidade participante* contendo o texto *SLTI*, e com o cadastramento de filtros para todas as unidades resultantes. Nota-se a diferença entre o uso do termo consulta, que indica um projeto de trabalho no sistema, e do termo busca, que indica uma pesquisa em cima das instâncias contidas em uma consulta. A Figura 56 mostra a busca realizada, ressaltando o campo de busca e a seleção de unidades para serem salvas como filtros. A busca mostra ainda o número de instâncias que passaram por cada unidade. Nessa etapa, a consulta foi reduzida para um total de 128.066 instâncias. Parte dos filtros resultantes é exibida na Figura 57.

Figura 56 – Busca por unidades participantes contendo o texto SLTI

Figura 57 – Filtros gerados pela busca por unidades participantes

Ao finalizar o primeiro passo, uma busca por *assuntos* foi necessária para se avaliar a estrutura dos tipos de processo presentes. A Figura 58 mostra o resultado dessa consulta. Nota-se que aproximadamente 65% dos processos não possuem assunto cadastrado. Os demais assuntos visualizados são muito genéricos, não refletindo processos específicos. Dessa forma, essa classificação de processos não permite a identificação de instâncias relacionadas para a etapa de descoberta de modelos. Nota-se que uma busca retorna como resultado somente as instâncias relacionadas à consulta atual. Embora os resultados da Figura 58 se restrinjam à SLTI, a base completa do sistema possui uma divisão entre assuntos similar.

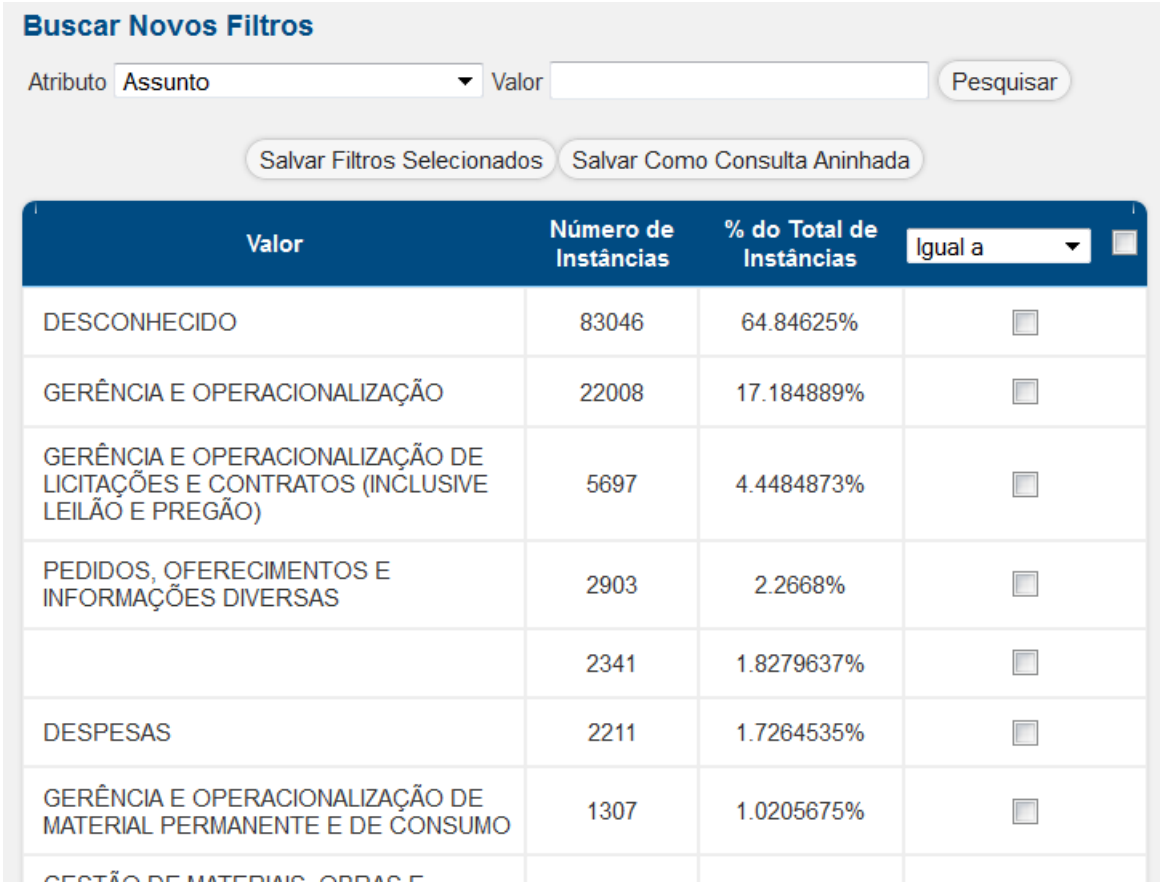

Figura 58 – Busca por assuntos

O atributo de *descrição*, por sua vez, embora seja cadastrado em campo textual livre, permite um entendimento muito maior a respeito das instâncias analisadas. A Figura 59 mostra uma busca por todas as descrições das instâncias contidas na consulta atual. Como o cadastro desse campo é em texto livre, existem diversas variações de uma mesma descrição, com alterações de, por exemplo, pontuação, plural, sinônimos, erros de digitação, dentre outros. Muitos usuários cadastram informações específicas de uma instância no campo de descrição, como, por exemplo, datas, códigos e nomes de interessados.

O *processo de desfazimento de equipamento de informática* foi selecionado para análise. Porém, para não se restringir somente às instâncias contendo exatamente esse texto, uma nova busca foi realizada, por todas as descrições contendo o texto *defazi%info*, sendo que o caracter % indica qualquer número de caracteres. Essa busca retornou resultados satisfatórios, com um grande número de descrições relacionadas. Ela é ilustrada na Figura 60. Todas as descrições retornadas foram incluídas como filtros, resultando em um total de 1.273 instâncias na consulta.

Valor	Número de Instâncias	% do Total de Instâncias	Igual a
SOLICITA SENHA DE ACESSO AO SIASG.	6229	4.8638983%	
REINTEGRACAO DE SERVIDOR	6015	4.696797%	
SOLICITA SENHA DE ACESSO AO SIASG	963	0.75195605%	
DESFAZIMENTO DE EQUIPAMENTO DE INFORMATICA.	918	0.7168179%	
ENCAMINHA TERMO DE RESPONSABILIDADE.	885	0.69104993%	
SOLICITA SENHA DE ACESSO AO SISTEMA SIASG.	860	0.67152876%	
REINTEGRACAO DE SERVIDOR.	695	0.5426889%	
SOLICITA CADASTRAMENTO DE ACESSO AO SIASG.	541	0.42243844%	

Figura 59 – Busca por descrições

Valor	Número de Instâncias	% do Total de Instâncias	Igual a
DESFAZIMENTO DE EQUIPAMENTO DE INFORMATICA.	918	0.7168179%	
DESFAZIMENTO DE MATERIAL DE INFORMATICA.	106	0.08276982%	
DESFAZIMENTO DE BENS DE INFORMATICA.	64	0.049974233%	
DESFAZIMENTO DE MATERIAIS DE INFORMATICA.	36	0.028110506%	
DESFAZIMENTO DE EQUIPAMENTOS DE INFORMATICA.	15	0.01171271%	

Figura 60 – Busca por descrições com o texto desfazi%info

As 1.273 instâncias da consulta resultante foram utilizadas para a fase seguinte de mineração. Foi utilizado o minerador de heurísticas com suas configurações padrão, resultando no modelo de processo apresentado na Figura 61. Pode ser visualizada uma grande quantidade de unidades, de enlaces em espaguete e de eventos de início e fim, dificultando o entendimento do processo. Nota-se, ainda, que os modelos gerados nesta prova de conceito são a interpretação da configuração padrão do minerador de heurísticas, que busca o

comportamento mais frequente dos processos. Ou seja, nem todos os fluxos existentes nas instâncias originais foram traduzidos em enlaces no modelo.

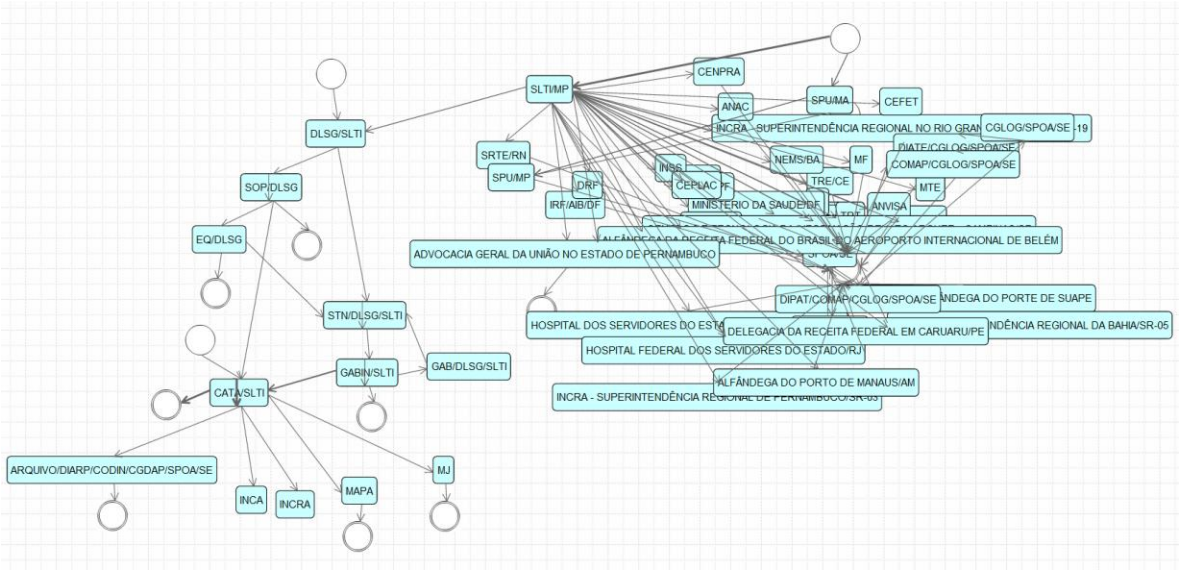

Figura 61 – Primeiro modelo de processo gerado para a consulta do CPROD

Para reduzir a quantidade de comportamentos pouco frequentes foram testadas duas abordagens: desabilitar a *heurística de todas as atividades conectadas*, o que reduziria a quantidade de unidades pouco frequentes do modelo, e a filtragem manual das unidades pouco frequentes. A segunda abordagem foi escolhida por ter resultado em um modelo de processo mais simples e por permitir uma maior obtenção de conhecimento a respeito do processo. Dessa forma, foram feitas buscas para três atributos: *unidades participantes*, *unidades iniciais* e *unidades finais*. A partir da análise da dispersão dos valores entre as instâncias, decidiu-se excluir os valores correspondentes a menos de 0,5% das instâncias. Ou seja, instâncias contendo unidades que participaram de menos de 0,5% do total de instâncias da consulta foram excluídas da análise, fazendo-se o mesmo para os dois outros atributos. Com isso, o modelo ilustrado pela Figura 62 foi obtido. Esse modelo é uma interpretação da configuração padrão do minerador de heurísticas, e não foi validado com as partes interessadas do processo. Embora a validação seja um passo importante para esta análise, ela está fora do escopo desta prova de conceito.

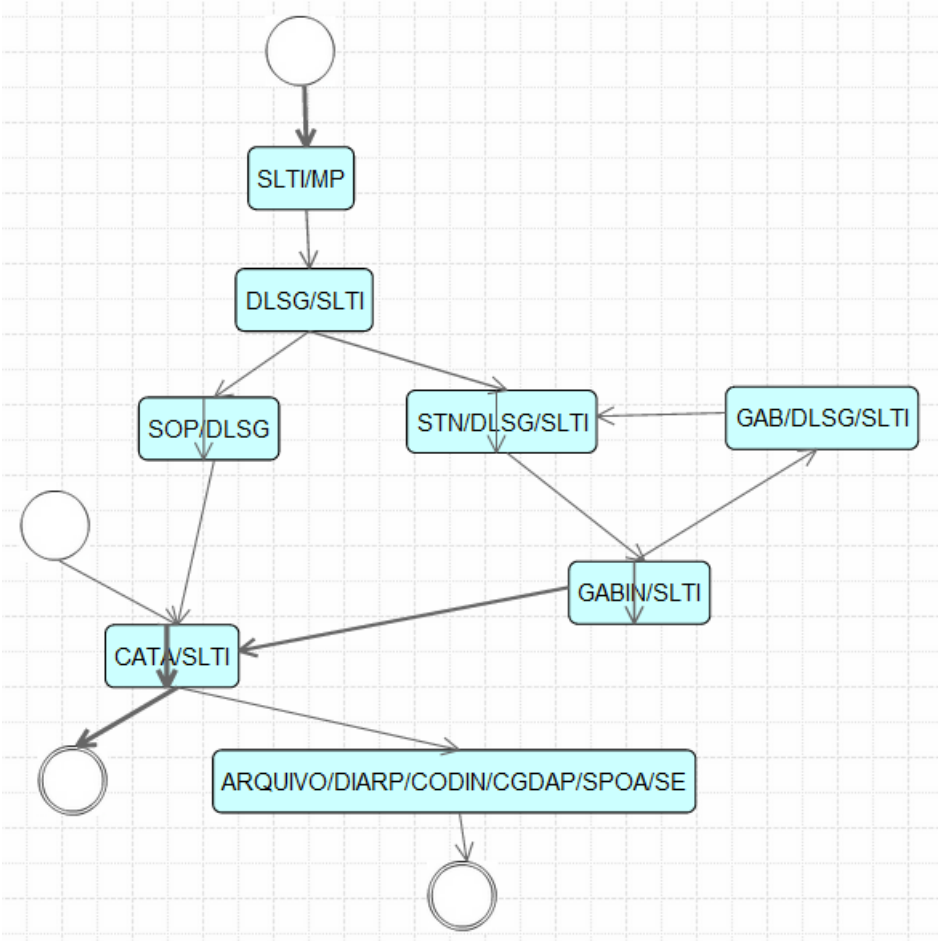

Figura 62 – Segundo modelo de processo gerado para a consulta do CPROD

A Figura 63 mostra a análise da consulta para o processo de desfazimento de equipamento de informática utilizando a animação de modelos. É possível perceber intuitivamente dois enlaces com um maior gargalo no fluxo das instâncias visualizadas, utilizando como base a cor das instâncias e a quantidade de instâncias exibidas em cada enlace. A instância em azul é a instância selecionada para a exibição de detalhes: é possível confirmar os atrasos relativos entre as unidades do fluxo utilizando essas informações. A análise do relatório de atraso por unidades, ilustrado na Figura 64, confirma a sugestão obtida pela análise da animação.

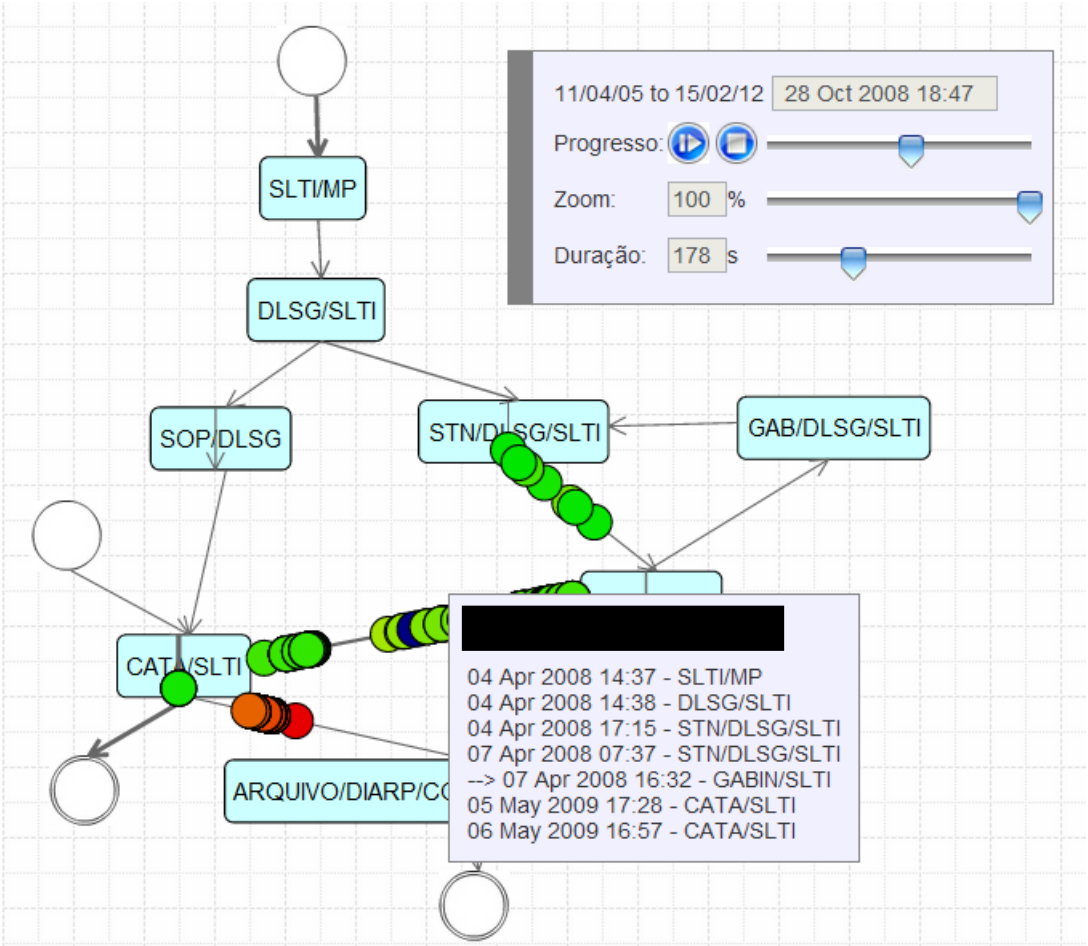

Figura 63 – Animação do modelo de processo gerado para o CPROD

Sigla	Instâncias	Trâmites Analisados	Trâmites Não Analisados	Atraso Mínimo	Atraso Médio	Atraso Máximo
CATA/SLTI	1210	1522	1070	0 dias	107 dias	1343 dias
GABIN/SLTI	938	1506	0	0 dias	96 dias	477 dias
STN/DLSG/SLTI	406	792	0	0 dias	6 dias	2253 dias
SOP/DLSG	271	669	0	0 dias	5 dias	31 dias
						343

Figura 64 – Relatório de desempenho para o CPROD

A análise realizada não foi em nenhum momento validada com as partes interessadas pelo processo; dessa forma, os resultados obtidos não podem ser conclusivos. Esta prova de

conceito, porém, mostra como existem casos em que é possível obter informações relevantes a respeito dos processos da organização utilizando o método MANA. Além disso, a reengenharia, uma etapa importante da abordagem proposta, através da qual os resultados estudados resultarão em melhorias reais, está fora do escopo desta prova de conceito. Dessa forma, os próximos passos para esta análise seriam:

- Reuniões com as partes interessadas pelo processo para discussão e validação do resultado.
- Ampliação do escopo da análise para outros processos.
- Melhoria da qualidade da classificação de processos da base de dados de origem.
- Alterar o sistema de informação original para ser ciente do fluxo dos processos que ele acompanha.

A profundidade da análise realizada nesta prova de conceito seria muito dificultada caso se utilizasse framework ProM. Não se pode assumir que o analista de processos saiba pesquisar diretamente um banco de dados utilizando consultas SQL para verificar os processos existentes e extrair arquivos contendo logs de eventos. Dessa forma, não seria possível explorar a base de dados para identificar os tipos de processos presentes e passíveis de mineração. Caso fossem utilizados os assuntos estruturados de processos para extrair diferentes logs de eventos, aproximadamente 2/3 dos processos da base ficariam de fora, dado que eles não possuem um assunto cadastrado. A utilização de algoritmos de clusterização, além de não se adequar a uma grande quantidade de comportamentos distintos (seria necessária uma enorme quantidade clusters), perderia informações importantes sobre a origem e o significado de cada cluster. A abordagem do minerador fuzzy, por sua vez, que agrega conjuntos de atividades em um modelo único, não se adequa às necessidades desta prova de conceito, cujo objetivo é separar tipos de processo individuais da base de origem.

Uma alternativa seria extrair um log de eventos para cada descrição das instâncias. Porém, mesmo com o escopo limitado à Secretaria de Logística e Tecnologia da Informação, a base de dados possui aproximadamente 84 mil descrições cadastradas. A análise utilizando o método MANA, por sua vez, agregou um total de 125 descrições relacionadas em uma consulta única para a identificação de um modelo de processo. Além disso, a integração das atividades de seleção de instâncias e de descoberta de modelos em uma única ferramenta

permite a adoção de uma abordagem exploratória, com a execução de testes sucessivos. A filtragem realizada pode ser adaptada iterativamente caso o modelo de processo gerado não seja considerado satisfatório.

Embora esta prova de conceito tenha se restringido somente à utilização de filtros manuais para um processo específico, e seus resultados não tenham sido validados com as partes interessadas, ele fornece indicativos de que existem casos em que é possível aprimorar a análise de mineração de processos em relação àquela possibilitada pelas ferramentas existentes na literatura. Especificamente, foi utilizado o método MANA, cuja seleção de instâncias se dá a partir de uma base padrão de instâncias de processo.

6.2 Sistema de Acompanhamento de Processos da UFRJ – SAP

O Sistema de Acompanhamento de Processos – SAP possibilita o registro e o acompanhamento de trâmites dos processos administrativos da Universidade Federal do Rio de Janeiro. Ele gerencia processos de registro de diploma, pagamento, aquisição, contratação, concessão de auxílios, dentre muitos outros. A base de dados desse sistema foi analisada nesta prova de conceito do ponto de vista da mineração de processos. No momento da análise, o sistema possuía aproximadamente 1,7 milhões de processos cadastrados, com 3,5 milhões de trâmites entre as unidades da universidade. Devido à grande quantidade de dados cadastrados, esta prova de conceito foi limitada ao estudo do processo de registro de diploma para a Escola Politécnica. Uma abordagem análoga, porém, pode ser utilizada para os demais processos da base.

6.2.1 Estrutura da Base de Dados

Esta seção tem como objetivo apresentar a base de dados do sistema SAP, indicando como seus conceitos estão relacionados à base padrão do método MANA. A modelagem de dados lógica da base é ilustrada pela Figura 65. O modelo foi simplificado para mostrar somente as informações que foram utilizadas; a base de dados completa possui um maior número de tabelas e atributos.

A entidade central da base é o *processo*. Um processo representa uma instância de processo cadastrada pelo sistema. Cada processo possui um *resumo* cadastrado em campo de texto livre, uma *data* de cadastramento, uma *unidade* e o *interessado* pelo processo (geralmente uma pessoa). Cada processo possui ainda um *assunto* dentre 242 pré-definidos.

Aproximadamente 250 mil processos foram cadastrados com o *assunto não informado*, ou 7% da base, uma proporção muito menor do que aquela encontrada na prova de conceito realizada com o sistema CPROD. Porém, apesar da quantidade de assuntos e da consistência de seu cadastramento, eles não identificam processos únicos. Alguns dos assuntos mais frequentes na base incluem *registro de diploma, revalidação/reconhecimento, assuntos acadêmicos, pagamento e aquisição*. O registro de diploma, por exemplo, inclui processos muito distintos entre si, como será analisado posteriormente.

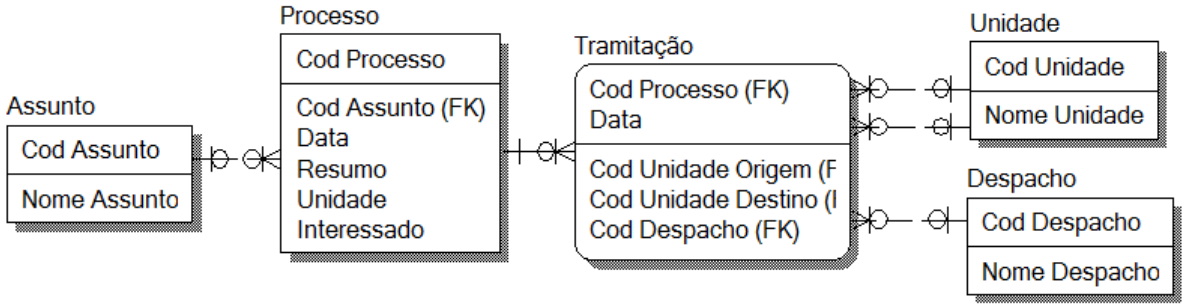

Figura 65 – Modelo de dados lógico simplificado do sistema SAP

A entidade tramitação registra cada vez que um processo foi repassado de uma *unidade de origem* a uma *unidade de destino*. Uma *unidade* pode ser interna ou externa à universidade. Um *despacho* é cadastrado, indicando, idealmente, a atividade a ser executada; porém, seus dados são muito genéricos. Exemplos de despacho incluem *providenciar/atender* e *comentar/opinar/analisar*. Como não foram obtidos bons resultados com a modelagem de processos utilizando esses dados, optou-se por modelar somente o fluxo entre unidades.

Tendo em vista os desafios encontrados, este trabalho optou por utilizar as instâncias de um mesmo assunto como ponto de partida para a mineração de processos. Esse conjunto de instâncias pode então ser refinado utilizando as demais informações cadastradas, como o resumo e as unidades participantes de uma instância. O detalhamento da análise realizada será descrito na seção seguinte. A Tabela 7 mostra como os dados encontrados na base de dados do SAP se relacionam aos conceitos do método MANA. Esses dados foram migrados através de uma carga ETL.

Tabela 7 - Mapeamento entre conceitos do SAP e do MANA

SAP	MANA
Processo	Instância de processo
Código do processo	Identificador da instância
Data	Data de início da instância
Resumo	Descrição da instância
Assunto	Assunto
Tramitação	Evento
Unidade do trâmite	Unidade
Unidade da instância	Origem
Interessado	Interessado

6.2.2 Mineração e Análise dos Processos

O fluxo de trabalho com a ferramenta MANA se inicia com a criação de uma consulta. Uma consulta começa contendo todas as instâncias da base de dados, que são reduzidas sucessivamente através de filtros. A Figura 66 mostra uma busca por todos os assuntos da base. Como dito anteriormente, nota-se a diferença entre o uso do termo consulta, que indica um projeto de trabalho no sistema, e do termo busca, que indica uma pesquisa em cima das instâncias contidas em uma consulta.

O assunto de registro de diploma foi escolhido para as análises realizadas no restante desta prova de conceito. A abordagem utilizada pode ser facilmente transferida para os demais processos da universidade. Este assunto foi selecionado por ser aquele com o maior número de registros na base, além de gerar frequentes reclamações dos alunos devido aos atrasos para a obtenção de seus diplomas após sua formatura. O registro de diploma é interessante do ponto de vista científico devido à complexidade para sua análise. Ele possui a participação de diversas unidades da universidade. Algumas unidades participam somente da emissão de diplomas específicos da sua escola ou faculdade, como, por exemplo, a Escola Politécnica ou a Faculdade de Letras. Porém, algumas unidades centrais da universidade, como a Divisão de Diplomas, participam do registro de todos os diplomas.

Figura 66 – Busca por assuntos

Com o filtro de assunto, a consulta passou a conter 362.422 instâncias. Porém, a descoberta de modelos utilizando somente esse filtro resultou em um processo em espaguete. Identificou-se que, além de diferentes escolas possuírem fluxos diferentes para o processo, grande parte das instâncias é na realidade relacionada a diplomas de universidades externas cadastrados na UFRJ para fins diversos. Numa tentativa de dividir a consulta em conjuntos menores de instâncias, o escopo do estudo foi reduzido aos processos da Escola Politécnica. Isso foi feito através da inclusão de um filtro com o atributo *unidade inicial* e o valor *Escola Politécnica*. Esse filtro foi utilizado porque existe uma abordagem análoga no framework ProM, analisada posteriormente nessa seção.

Uma inspeção do filtro de *ano*, porém, mostrou que a grande maioria dos processos filtrados corresponde a instâncias iniciadas a partir do ano de 2005. A existência de poucas instâncias para datas muito antigas é esperado, mas não para datas recentes. O filtro de ano inclui uma análise gráfica, indicando os anos no eixo x e o número de instâncias no eixo y. O primeiro gráfico da Figura 67 mostra esse resultado. Numa tentativa de solucionar o problema encontrado, foi feita uma busca pela *origem* das instâncias. Nota-se que a origem de uma instância não representa necessariamente a unidade que executou a primeira atividade do

processo. Dessa forma, identificou-se que grande parte das instâncias com origem na Escola Politécnica tem seu fluxo de atividades iniciado, na realidade, na Pró-Reitoria de Graduação – PR1.

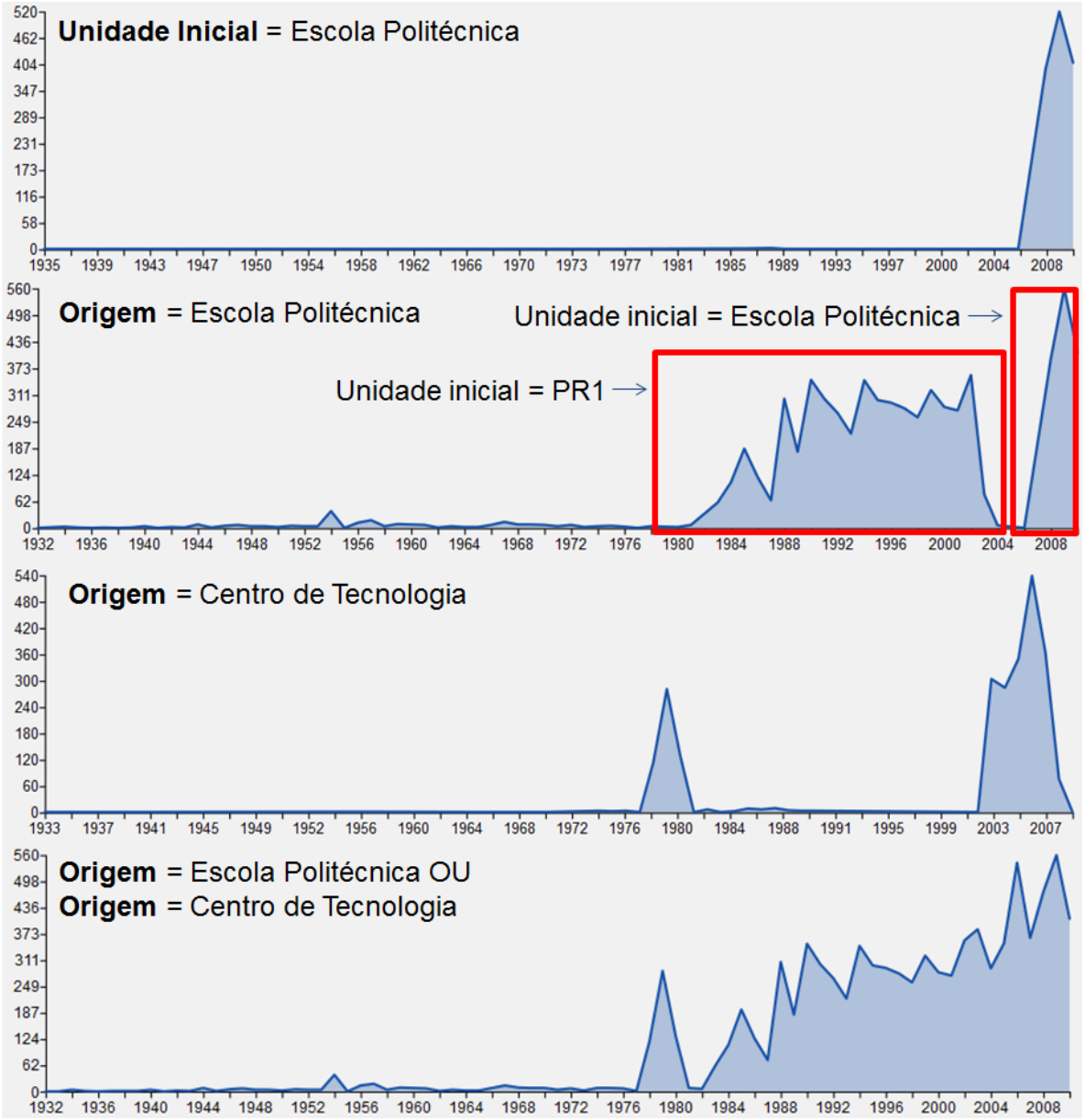

Figura 67 – Número de instâncias por ano para diferentes filtros

Mesmo com a utilização do filtro de origem, identificou-se um período recente no gráfico com uma quantidade muito reduzida de instâncias cadastradas. Identificou-se que uma grande quantidade de instâncias está cadastrada sob a origem *Centro de Tecnologia*, para os anos faltantes. O último gráfico da Figura 67 mostra a união dessas duas origens, que foi

utilizada nesta prova de conceito. Embora a análise gráfica forneça fortes indícios de que a união realizada indique o mesmo processo, isso não foi validado com as partes interessadas do processo, e as causas dessa diferenciação entre as origens das instâncias não foi analisada com maiores detalhes. A filtragem pelas origens resultou em um total de 9.109 instâncias na consulta.

A próxima etapa realizada envolveu a análise das instâncias que tiveram eventos (trâmites) registrados. Foi identificado que somente instâncias iniciadas a partir do ano de 1997 possuem consistentemente eventos. A Figura 68 mostra essa análise, contando o número de instâncias com eventos para cada ano considerado. Para eliminar os dados de anos anteriores, foram filtradas somente as instâncias iniciadas a partir do ano de 1997, resultando em 5.119 instâncias.

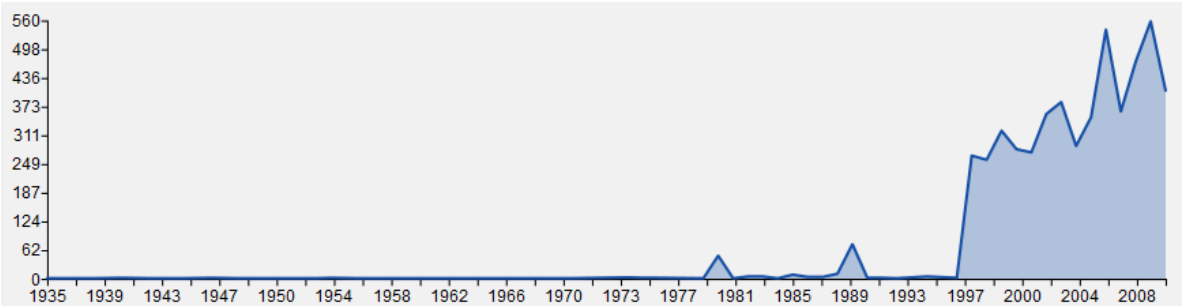

Figura 68 – Instâncias com eventos registrados

Outra etapa de filtragem exploratória envolveu a análise das *descrições* das instâncias. Descrições diversas foram identificadas. Alguns exemplos incluem:

- SOLICITAÇÃO DE REGISTRO DE DIPLOMA.
- SOLICITAÇÃO DE 2ª VIA DE DIPLOMA.
- ANEXO 01 DIPLOMA. (2A.VIA).
- REVALIDACAO DE DIPLOMA DE MESTRE

Para eliminar a grande disparidade de processos cadastrados sob o assunto registro de diploma, foram filtradas manualmente somente descrições consideradas como registros iniciais de diplomas de graduação. Outros filtros realizados incluem a remoção de *unidades finais* pouco frequentes (último trâmite em menos de 2% das instâncias, indicando provavelmente instâncias incompletas) e *unidades participantes* pouco frequentes

(participando de menos de 2% das instâncias, indicando provavelmente *outliers*). Os valores escolhidos para a filtragem de cada atributo são dependentes do entendimento do analista a respeito dos processos estudados, característica da abordagem exploratória proposta neste trabalho.

Ao final da fase de identificação de instâncias relacionadas, as 2.565 instâncias resultantes foram mineradas utilizando o minerador de heurísticas em sua configuração padrão. Considerando os principais filtros realizados, o modelo de processo representa os principais fluxos para o processo que foi chamado de *registro de diplomas de graduação de primeira via para a Escola Politécnica*. O modelo resultante é mostrado na Figura 69. Como foi modelado o fluxo entre unidades e não entre atividades, esse processo apresenta *loops*. A análise do processo com suas partes interessadas pode ser utilizada posteriormente para substituir cada unidade do modelo pelas atividades que ela executa, mitigando esse problema.

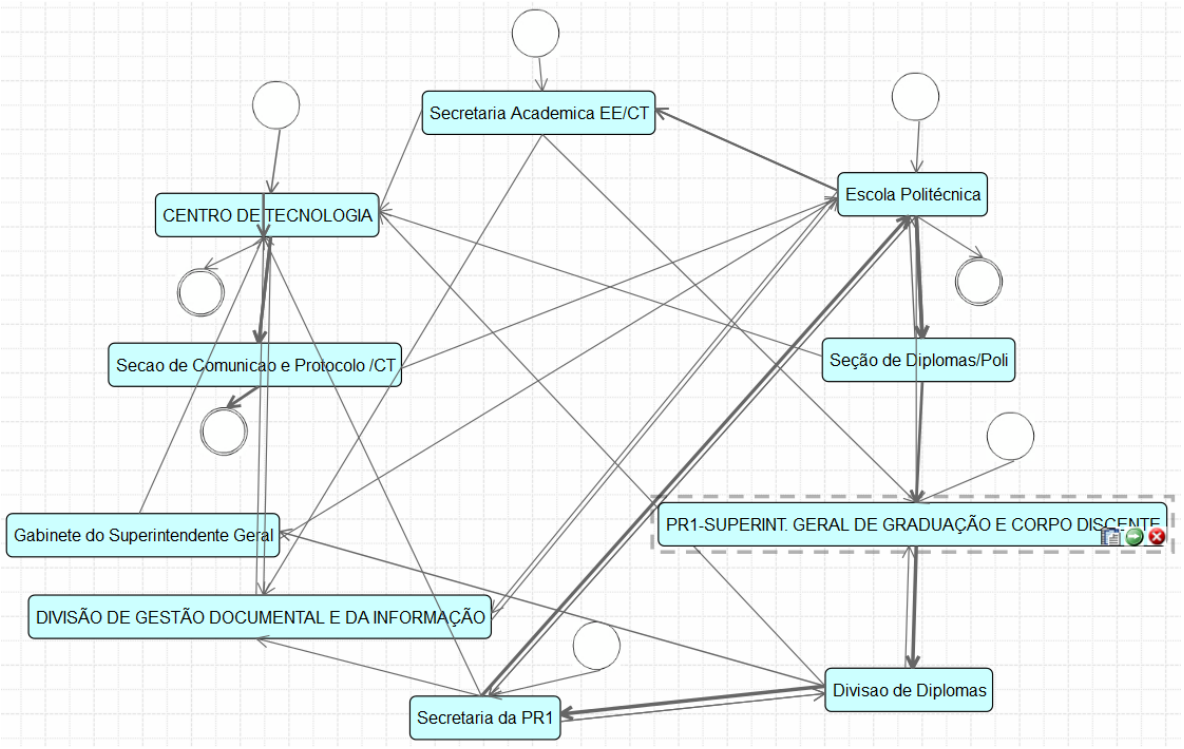

Figura 69 – Modelo do processo de registro de diplomas de graduação de primeira via para a Escola Politécnica

Para simplificar o processo modelado, foi utilizada a técnica de clusterização DWS discutida na seção 3.3.1. O algoritmo foi configurado para 2 clusters. Os modelos de processo resultantes para cada cluster são apresentados na Figura 70 e na Figura 71. Essa separação foi

considerada razoável para a análise atual; porém, o processo pode ser dividido em um maior número de clusters caso necessário.

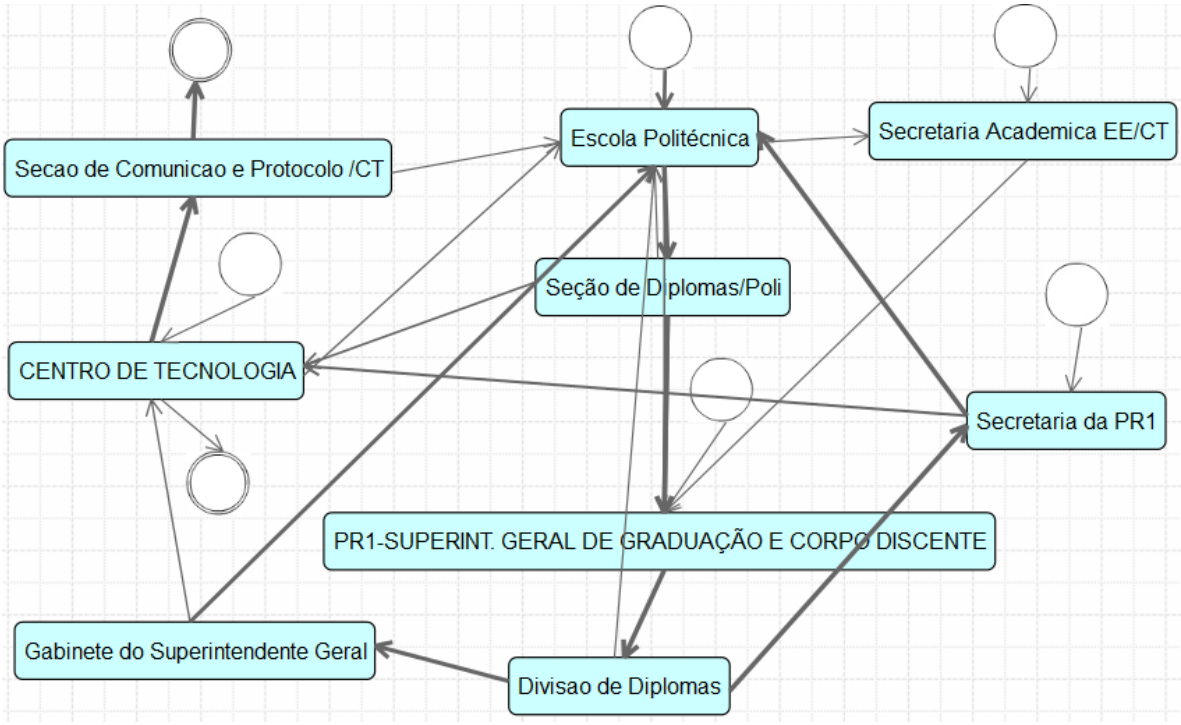

Figura 70 – Primeiro cluster para a consulta analisada

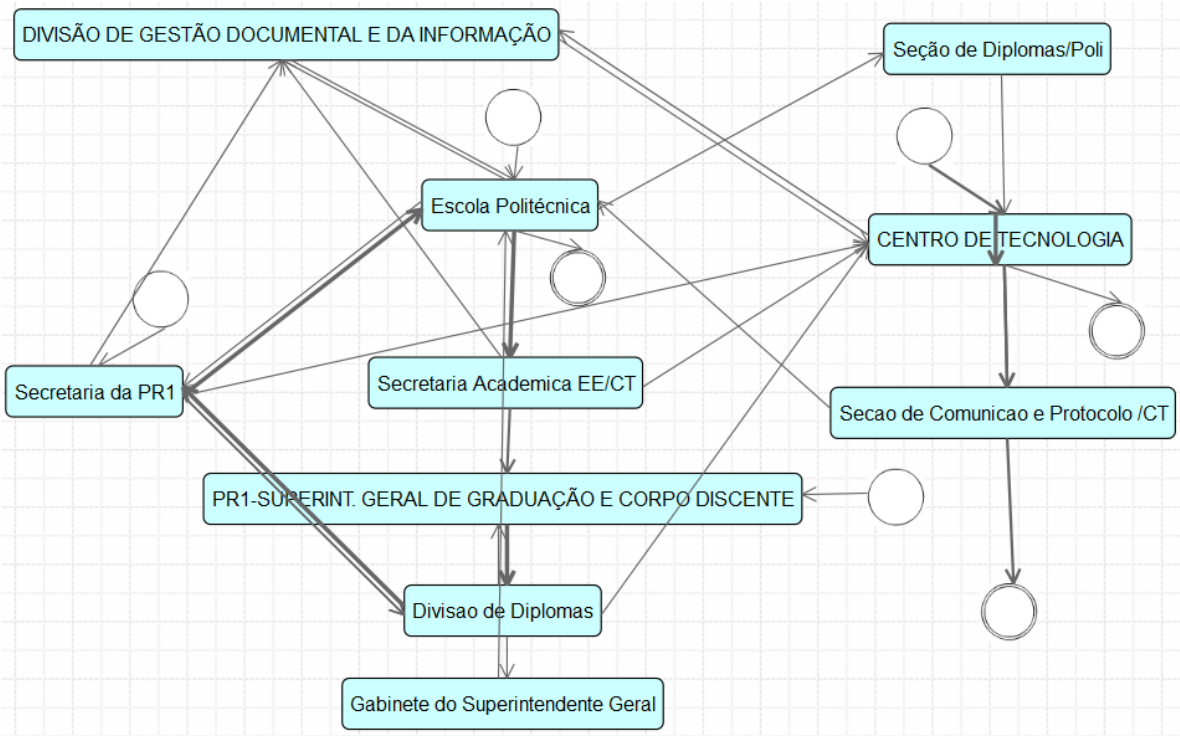

Figura 71 – Segundo cluster para a consulta analisada

O modelo de processo referente ao primeiro cluster foi analisado utilizando a animação de instâncias. Dessa forma, é possível analisar as deficiências de processo e entender o fluxo de instâncias em maiores detalhes. Pode-se observar uma grande quantidade de instâncias de atividades lentas com origem na Seção de Comunicação e Protocolo/CT para a Escola Politécnica, com duração de vários anos. Esse caso pode ser analisado em maiores detalhes, podendo indicar uma deficiência no processo, uma reabertura da instância ou a falta de uma atividade de armazenamento por falta de arquivista.

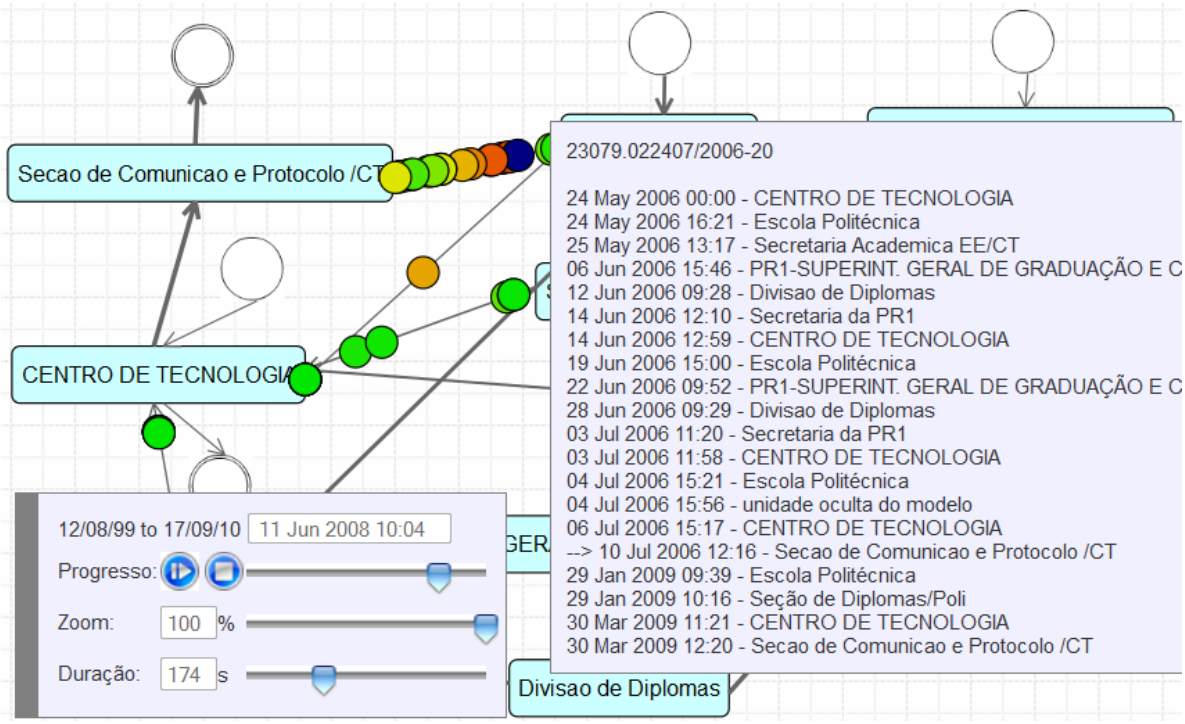

Figura 72 – Animação de instâncias para o primeiro cluster da consulta analisada

A Figura 73 e a Figura 74 mostram outras análises possíveis para a consulta. O relatório de desempenho de unidades permite avaliar o atraso mínimo, médio e máximo de cada unidade para as instâncias analisadas. O gráfico de linha do tempo, por sua vez, permite analisar a data de chegada e de saída de cada instância de uma unidade. O eixo x indica a linha do tempo, enquanto o eixo y representa somente uma sequência de instâncias. Cada reta mostra o tempo que a instância passou na unidade. É possível verificar como muitas instâncias são resolvidas de uma só vez no processo.

Sigla	Instâncias	Trâmites Analisados	Trâmites Não Analisados	Atraso Mínimo	Atraso Médio	Atraso Máximo	Dispersão	Timeline
Secao de Comunicao e Protocolo /CT	1177	55	1176	0 dias	347 dias	1296 dias	Dispersão	Timeline
Secretaria Academica EE/CT	72	96	0	0 dias	23 dias	333 dias	Dispersão	Timeline
Seção de Diplomas/Poli	1199	3204	0	0 dias	21 dias	383 dias	Dispersão	Timeline
Divisao de Diplomas	1199	2583	0	0 dias	9 dias	82 dias	Dispersão	Timeline
Escola Politécnica	1200	3810	0	0 dias	6 dias	3407 dias	Dispersão	Timeline
Gabinete do								

Figura 73 – Relatório de desempenho para o primeiro cluster da consulta analisada

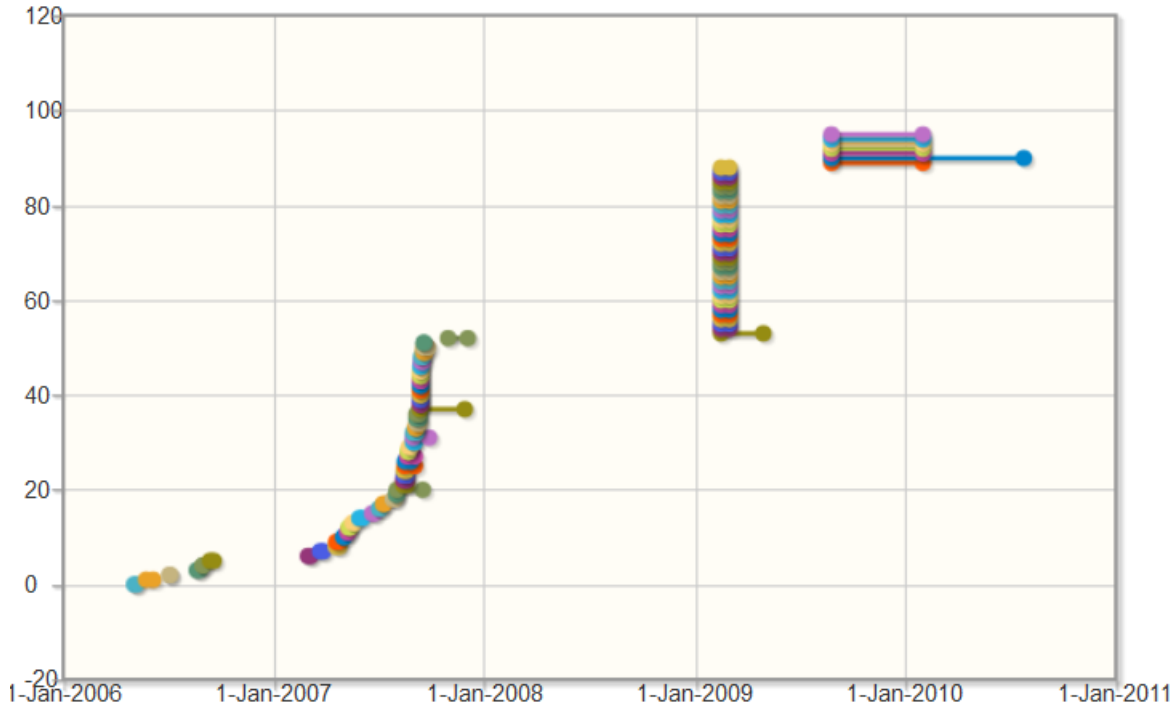

Figura 74 – Gráfico de linha do tempo para a Secretaria Acadêmica EE/CT

A segunda parte desta prova de conceito envolve a análise dos mesmos dados utilizando o framework ProM 6.1. As instâncias relacionadas ao assunto de registro de diploma foram exportadas para um arquivo MXML e carregadas para a ferramenta. Uma tentativa inicial de minerar o processo utilizando o minerador de heurísticas foi feita. Parte do modelo resultante é ilustrado pela Figura 75, exibindo a grande disparidade entre as instâncias carregadas. O minerador fuzzy retornou um modelo de processo similar, não tendo sido capaz de lidar com a grande quantidade de instâncias carregadas.

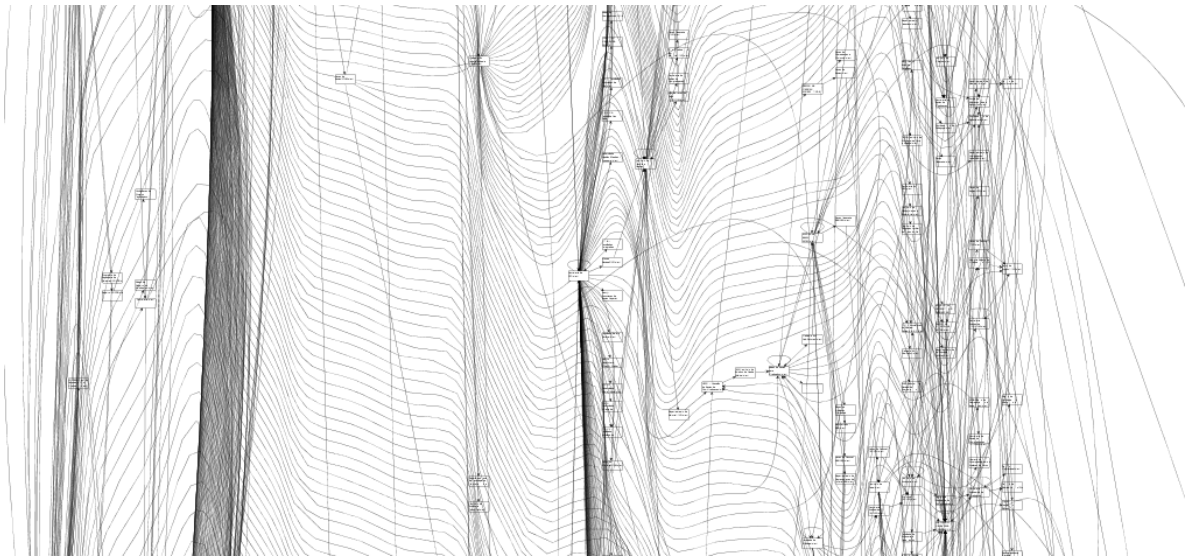

Figura 75 – Trecho do processo de registro de diploma gerado com o minerador de heurísticas

Para reduzir a complexidade do modelo, foram utilizados os filtros disponíveis pelo framework: *eventos de início, eventos de fim e eventos*. O framework ProM permite filtrar somente os casos mais frequentes para cada filtro disponível. Essa análise, porém, peca pelo fato de retornar um modelo de processo cujo resultado é de difícil interpretação, visto que ele procura simplificar uma grande quantidade de tipos de processo em um só modelo. A análise usando o método MANA, por sua vez, permite a obtenção de um maior nível de conhecimento a respeito do modelo de processo gerado, através da etapa de identificação de instâncias relacionadas.

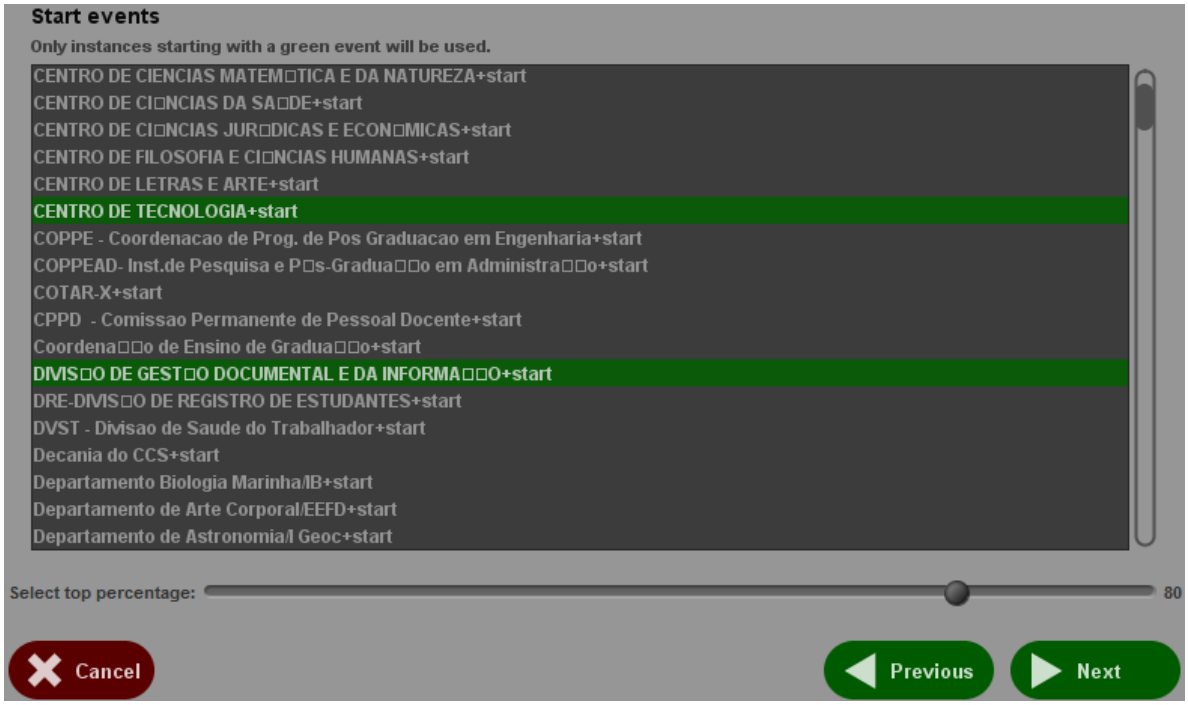

Figura 76 – Filtragem de eventos com o framework ProM 6.1

Para limitar o escopo à Escola Politécnica, uma abordagem seria utilizar o filtro de eventos de início. Um modelo de processo obtido dessa forma é exibido na Figura 77. Porém, como estudado anteriormente, essa abordagem é falha, retornando apenas instâncias iniciadas a partir do ano de 2005. A abordagem correta seria incluir instâncias com eventos de início relacionados à Escola Politécnica, ao Centro de Tecnologia, à PR1 e à Secretaria da PR1. Porém, como a pró-reitoria atende toda a universidade, o escopo não seria limitado somente à Escola desejada. Dessa forma, a análise realizada com o método MANA não pode ser realizada diretamente com o framework ProM.

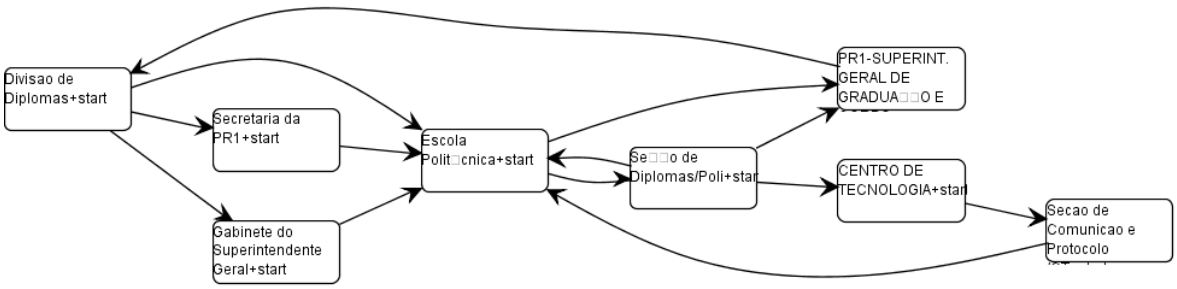

Figura 77 – Modelo de processo após filtragem com o framework ProM

Vale ressaltar que a versão 5.2 do ProM possui mais filtros disponíveis do que a versão 6.1. Porém, esses filtros avançados exigem que o usuário saiba previamente os valores que deseja filtrar, não permitindo a exploração dos valores disponíveis. O filtro de *atributo*, por exemplo, permite a filtragem específica de um valor dado para um atributo dado. Esse filtro é ilustrado pela Figura 78. Porém, sua utilização não é viável em situações reais como as estudadas neste trabalho, onde a exploração dos valores existentes é importante para a seleção de instâncias relacionadas. Além disso, a utilização direta de abordagens de clusterização automática, também presentes na versão 5.2, não foi capaz de retornar modelos de processo de complexidade considerada razoável.

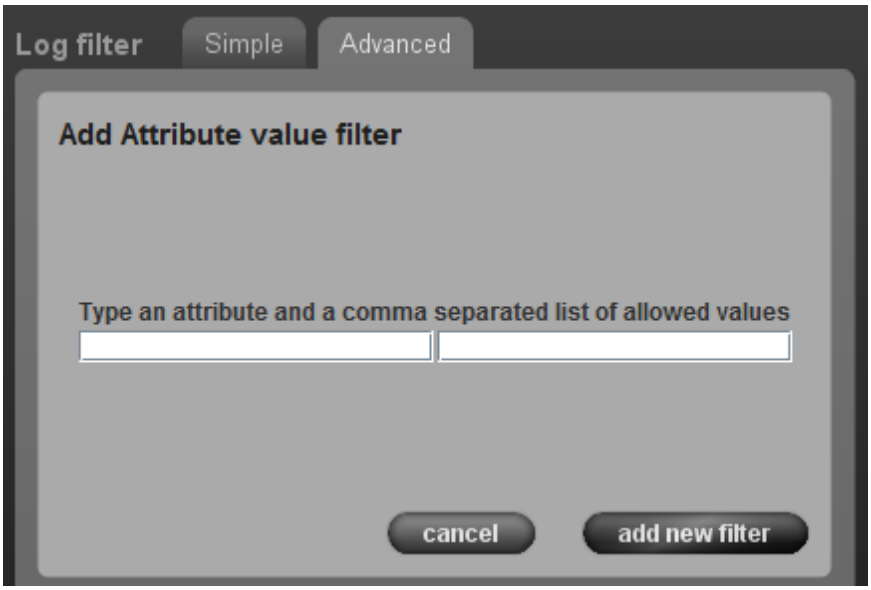

Figura 78 – Filtro de atributo do framework ProM 5.2

6.3 Considerações finais

Este capítulo apresentou duas provas de conceito que fornecem indícios de como a aplicação do método proposto pode aprimorar a mineração de processos desestruturados em algumas situações. Os estudos realizados foram limitados em escopo, analisando somente um processo de cada base de dados e não dando continuidade à validação dos resultados com as partes interessadas pelos processos e sua reengenharia. O capítulo seguinte apresenta as conclusões finais deste trabalho e suas direções futuras.

Capítulo 7 – Conclusões

Este capítulo apresenta as conclusões deste trabalho. A seção 7.1 descreve as considerações finais, ressaltando os principais pontos do trabalho desenvolvido e como ele se insere no contexto do gerenciamento de processos de negócio. Os principais resultados atingidos e contribuições são indicados na seção 7.2, enquanto a seção 7.3 ressalta as limitações da pesquisa realizada. Finalmente, a seção 7.4 apresenta os trabalhos futuros, envolvendo a evolução da abordagem proposta e o aprofundamento das provas de conceito apresentadas.

7.1 Considerações Finais

O gerenciamento de processo de negócio é importante para organizações que buscam o autoconhecimento e a execução de suas atividades de maneira consistente e otimizada. Inserida nesse contexto está a mineração de processos, área de pesquisa recente que possui como um de seus objetivos principais o tratamento de deficiências existentes na modelagem de processos tradicional. Ao utilizar dados reais de instâncias de processos, é possível gerar modelos que traduzam sua execução real, permitindo a contestação de visões tendenciosas das partes interessadas pelo processo. Além disso, a mineração de processos fornece um ponto de partida para as atividades de modelagem, reduzindo a quantidade de recursos exigidos. Seus resultados podem motivar o início de atividades de reengenharia organizacional, que do contrário seriam dificilmente alavancadas em algumas organizações. Em especial, organizações públicas, que motivaram a execução deste trabalho, têm muito a se beneficiar da utilização de técnicas de mineração que forneçam um ponto de partida para projetos de modelagem de processos.

A mineração de processos desestruturados, porém, apresenta diversos desafios para as técnicas de mineração de processos existentes atualmente. Diversas abordagens têm sido propostas na literatura para lidar com esse tipo de processo. Existem casos, porém, em que as ferramentas existentes que suportam essas técnicas não são capazes de retornar resultados satisfatórios. Isso ocorre principalmente quando existe dificuldade em separar um conjunto de instâncias de processo de um mesmo tipo, ponto de partida para os algoritmos de clusterização e de descoberta de modelos. Informações importantes para a seleção de instâncias são geralmente perdidas, e poderiam ser utilizadas através de uma abordagem

exploratória que agregue conhecimento à identificação de processos na base de dados analisada. Essa abordagem pode ser utilizada mesmo quando a base de origem possui uma grande quantidade de instâncias, como mostrado no capítulo 6. A exploração da base deveria ser integrada em um fluxo de trabalho conjuntamente às atividades de mineração de processos, aumentando a possibilidade de extrair conhecimento a partir da análise dos processos e permitindo que analistas de negócio utilizem técnicas de mineração.

Dessa forma, este trabalho introduziu o método MANA, definindo um fluxo de identificação, mineração, análise e reengenharia de processos tendo como ponto central a utilização de uma base de dados padrão de instâncias. Para cada etapa da abordagem, foi proposto um conjunto de atividades, envolvendo a análise incremental da base dados para a obtenção de conhecimento a respeito dos processos da organização. A identificação de processos envolve a exploração da base, com a utilização de buscas sobre os atributos da base, além de técnicas automatizadas de clusterização e de busca por instâncias similares. O conjunto de instâncias relacionadas identificado deve então ser iterativamente analisado com algoritmos de mineração e refinado, até a obtenção de um modelo considerado razoável pelo analista e validado com suas partes interessadas. O processo deve então ser analisado, utilizando, principalmente, técnicas de visualização que permitam a identificação de deficiências e a motivação para a necessidade de atividades de reengenharia. Estas, por sua vez, envolvem a construção de modelos de processo *to-be*, a melhoria da qualidade de dados da base de origem, com uma melhor tipificação de processos, e a adoção de um sistema de informação que seja ciente dos processos executados. A maioria das atividades descritas pelo método foi implementada em uma ferramenta, cujo objetivo é viabilizar a utilização do fluxo de trabalho proposto.

Duas provas de conceito foram apresentadas neste trabalho, utilizando dados extraídos de sistemas de informação de duas organizações públicas: a Universidade Federal do Rio de Janeiro e o Ministério do Planejamento, Orçamento e Gestão. Eles tiveram como objetivo principal mostrar que existem casos em que a abordagem proposta é capaz de extrair um maior nível de conhecimento e modelos de processo mais precisos do que as abordagens existentes na literatura. Os sistemas analisados possuem em comum o fato de acompanharem o trâmite de processos entre unidades dessas organizações, permitindo a entrada de dados em campos de texto livre e sem sugerir algum fluxo de atividades para os processos executados.

7.2 Resultados e Contribuições

O principal objetivo deste trabalho, introduzido na seção 1.3, está relacionado à definição de um método com um fluxo de atividades de identificação, mineração, análise e reengenharia de processos desestruturados, utilizando uma base de dados padrão de instâncias de processos. Objetivos correlatos envolveram a definição de uma ferramenta para suportar o método, o foco na identificação de processos com a exploração da base, facilitar o acesso a técnicas de mineração de processos e a motivação da reengenharia de processos nas organizações. Esses objetivos foram alcançados pela definição do método MANA, descrita no capítulo 5, e o desenvolvimento de uma ferramenta que permita a utilização do fluxo de trabalho proposto.

Os principais resultados e contribuições deste trabalho incluem:

- Identificação da dificuldade em se minerar processos desestruturados utilizando as técnicas existentes na literatura, em casos onde o sistema de informação de origem permite a entrada de dados textual em campo livre e não utiliza uma classificação que permita a separação de conjuntos de instâncias relacionadas;
- Levantamento da hipótese relacionada à possibilidade de melhorar o tratamento do problema, em relação às abordagens existentes, utilizando uma base de dados padrão como centro da identificação de instâncias relacionadas;
- Proposição do método MANA, que define um fluxo para a identificação, mineração, análise e reengenharia de processos desestruturados;
- Foco na atividade de seleção de instâncias de processo relacionadas, utilizando filtros sobre atributos baseados em informações presentes na base padrão, o que permite a obtenção incremental de conhecimento a respeito dos processos analisados e a obtenção de modelos de processo mais precisos;
- Desenvolvimento de uma ferramenta para suportar o método proposto, permitindo a utilização do fluxo proposto e facilitando o acesso a técnicas de mineração de processos;
- Explicitar a importância da etapa de reengenharia, envolvendo, além da construção de modelos de processo *to-be*, a reestruturação dos tipos de processo na base de dados

original e do sistema de informação que suporte esses processos. Essas atividades são importantes para corrigir a raiz dos problemas identificados nos sistemas com as características particulares que motivaram este trabalho. A utilização de análises visuais, como a animação de processos, contribui para exibir o grau atual de desestruturação dos processos de uma organização e auxiliar no desenvolvimento de projetos de reengenharia;

- Descrição de duas provas de conceito, fornecendo exemplos de sistemas de informação contendo as características estudadas neste trabalho, e mostrando como a abordagem proposta pode contribuir para a melhoria da análise dos processos apresentados.

Resultados obtidos no decorrer deste trabalho foram publicados em conferências, incluindo:

- Esposito et al. (2011): Primeira abordagem utilizada estudando a mineração do processo de registro de diplomas de uma universidade. Apresenta o problema de mineração de processos com dados extraídos de sistemas de protocolo de organizações públicas. Após uma filtragem inicial, foram utilizados dois níveis de clusterização para separar as instâncias mineradas entre as diferentes escolas da universidade e simplificar os modelos de processo obtidos. Nota-se que a prova de conceito apresentada no capítulo 6 deste trabalho expande essa abordagem, através a exploração da base padrão, permitindo a obtenção de resultados mais precisos do que proposto no artigo.
- Esposito et al. (2012): Apresenta uma visão geral da abordagem MANA e de como ela procura lidar com a mineração de processos desestruturados, ressaltando suas principais contribuições em relação à abordagem do framework ProM. É apresentado um exemplo para mostrar a aplicação da abordagem. Esse artigo foi aceito para publicação em setembro de 2012 no *Workshop Business Process Intelligence* (BPI), a ser realizado em conjunto à conferência BPM 2012.

7.3 Limitações

O método MANA e a ferramenta desenvolvida não se propõem a substituir as abordagens existentes de mineração de processos. O framework ProM, por exemplo, suporta

uma grande quantidade de técnicas que estão fora do escopo deste trabalho. Ele possui muitos algoritmos de descoberta de modelos que não foram utilizados pela ferramenta desenvolvida. Este trabalho indica que, nas situações específicas estudadas, a integração de atividades de seleção de instâncias de uma base padrão e uma melhor definição do fluxo de trabalho pode apresentar melhores resultados em alguns casos de aplicação da mineração de processos.

O método MANA propõe a utilização de uma base de dados padrão de instâncias de processo. Porém, as entidades e atributos propostos surgiram apenas da análise das particularidades das bases de dados utilizadas para as provas de conceito. Dessa forma, não foi definido o formato ideal de uma base de dados padrão de processos.

O módulo de filtros da ferramenta desenvolvida, embora permita a busca exploratória, age de maneira mais passiva do que ativa. Quando o usuário realiza uma busca para um atributo, sua sugestão de valores se limita à ordenação dos resultados a partir do número de instâncias que possui cada valor retornado. Melhorias de interface devem ser investigadas, com base principalmente em abordagens de busca facetada (Hearst 2006), para sugerir filtros ao usuário e exibir simultaneamente valores relacionados a diversas dimensões da instância.

A abordagem de análise de processos utilizando relatórios e a animação de instâncias é limitada. O foco na análise de uma animação foi escolhido por sobressaírem aspectos intuitivos de desempenho e pelo impacto visual causado, importante para motivar projetos de reengenharia para a alta gerência das organizações. Estudos posteriores são necessários para aprimorar a etapa de análise proposta pela abordagem.

Embora as provas de conceito apresentadas no capítulo 6 mostrem indícios de que é possível aprimorar a mineração de processos desestruturados utilizando o método proposto, em comparação às abordagens existentes, estudos posteriores são necessários para provar a hipótese apresentada. As provas de conceito indicaram que uma seleção de instâncias de processo mais específica foi possível. Porém, a validação dos resultados com as partes interessadas dos processos estudados, principalmente dos modelos gerados, não foi possível de ser realizada devido a limitações de tempo para a conclusão deste trabalho. Uma análise completa dos processos das bases de dados estudadas também foi dificultada pela grande quantidade de instâncias não relacionadas presentes. Nota-se ainda que, embora a descoberta

de modelos de processo retorne resultados baseados em dados reais, eles estão sujeitos à efetividade limitada do algoritmo utilizado e de suas parametrizações.

As atividades de reengenharia de processos são de extrema importância e representam os resultados reais da mineração de processos. A melhoria dos processos da organização e do sistema de informação que os suporta são o objetivo final do esforço realizado pela aplicação do método MANA. Porém, devido à complexidade de sua execução e de limitações de tempo, elas se encontram fora do escopo das provas de conceito realizadas neste trabalho.

7.4 Trabalhos Futuros

Trabalhos futuros envolvem, principalmente, o tratamento de situações que ficaram fora do escopo deste trabalho, descritas na seção 7.3. Em relação à abordagem desenvolvida, estudos posteriores são necessários para validar seu desempenho nas situações propostas. As provas de conceito desenvolvidas devem ter seu escopo ampliado, tanto no número de processos analisados quanto na continuidade das atividades de reengenharia de processos. Para isso, devem ser realizadas reuniões com as partes interessadas pelos processos para a discussão dos resultados encontrados, sua validação e a correção de pontos necessários na abordagem empregada. Projetos para aplicar uma tipificação estruturada de processos e adotar um sistema de informação ciente de processos são necessários nas duas situações estudadas. Vale ressaltar, porém, a complexidade dessa abordagem, dado o grande escopo dos sistemas de informação avaliados.

A ferramenta desenvolvida para suportar o método MANA pode ser aprimorada de diversas formas. Um aspecto seria ampliação dos dados suportados pela base padrão, e a correspondente disponibilização de novos atributos de filtragem que utilizem esses dados, possivelmente customizados pelo próprio analista. O módulo de filtros, que utiliza esses dados para a exploração do usuário, poderia ter sua interface aprimorada utilizando padrões comuns a ferramentas de busca facetada. Algumas considerações para esse tipo de interface são encontradas em Hearst (2006). Além disso, a carga de dados ETL para a base de dados padrão é feita por um analista de TI com conhecimento técnico; o suporte ferramental para a carga de dados eliminaria essa necessidade.

A implementação de novas técnicas de clusterização, inclusão de instâncias similares e de descoberta de modelos também é importante para aprimorar os resultados obtidos pela

utilização da ferramenta. Por exemplo, a utilização de informações semânticas permitiria aprimorar a identificação de instâncias similares. Novos métodos de análise também são necessários, como, por exemplo, a implementação de *dashboards* gerenciais baseados em *Key Performance Indicators*. A utilização de um armazém de dados de processos, como o proposto por Casati et al. (2007), é uma das alternativas futuras para viabilizar a inclusão de novas técnicas de análise a partir das instâncias filtradas por uma consulta de processo.

Referências Bibliográficas

van der Aalst, W. M. ., Weijters, A. J. M. ., Maruster, L., (2004), "Workflow Mining: Discovering process models from event logs", *IEEE Transactions on Knowledge and Data Engineering*, v. 16, n. 9, p. 1128-1142.

van der Aalst, W. M. P., (2004), "Business Process Management Demystified : A Tutorial on Models , Systems and Standards for Workflow Management", *Lectures on Concurrency and Petri Nets*, v. 3098, n. 3098, p. 1-65.

van der Aalst, W. M. P., (2011), *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. 1st Edition. ed. Springer.

van der Aalst, W. M. P., van Dongen, B. F., Herbst, J., Maruster, L., Schimm, G., Weijters, A. J. M. M., (2003a), "Workflow mining: a survey of issues and approaches", *Data & Knowledge Engineering*, v. 47 (nov.), p. 237–267.

van der Aalst, W. M. P., Gunther, C. W., (2007), "Finding Structure in Unstructured Processes: The Case for Process Mining". In: *Proceedings of the Seventh International Conference on Application of Concurrency to System Design*, p. 3–12, Washington, DC, USA.

van der Aalst, W. M. P., Hofstede, A. H. M., Weske, M., (2003b), "Business Process Management: A Survey", In: van der Aalst, W. M. P., Weske, M. [orgs.] (eds), *Business Process Management*, , chapter 2678, Berlin, Heidelberg: Springer Berlin Heidelberg, p. 1-12.

van der Aalst, W. M. P., Weijters, A. J. M. M., (2004), "Process mining: a research agenda", *Computers in Industry*, v. 53, n. 3 (abr.), p. 231-244.

Baranovskiy, D., (2012). Raphaël—JavaScript Library. Disponível em: http://raphaeljs.com/. Acesso em: 20 mar 2012.

Baureis, R., (2010). Basic rules of EPC modelling - ARIS BPM Community. Disponível em: http://www.ariscommunity.com/users/rbaureis/2010-03-22-basic-rules-epc-modelling. Acesso em: 7 fev 2012.

Ben-Yitzhak, O., Golbandi, N., Har'El, N., Lempel, R., Neumann, A., Ofek-Koifman, S., Sheinwald, D., Shekita, E., Sznajder, B., et al., (2008), "Beyond basic faceted search". In:

Proceedings of the international conference on Web search and web data mining, p. 33–44, New York, NY, USA.

Bose, R. P. J. C., van der Aalst, W. M. P., (2009), "Context Aware Trace Clustering: Towards Improving Process Mining Results". In: *Proceedings of SDM'2009*, p. 401-412

Brasil, (2008), "Instrução Normativa n° 04 de 19 de maio de 2008", *Diário Oficial da União*, v. 95, sec. 1, p. 95-97.

Bujis, J. C. A. ., (2010), *Mapping Data Sources to XES in a Generic Way*. Dissertação de Mestrado, Eindhoven University of Technology. Disponível em: http://www.processmining.org/_media/xesame/xesma_thesis_final.pdf.

Casati, F., Castellanos, M., Dayal, U., Salazar, N., (2007), "A generic solution for warehousing business process data". In: *Proceedings of the 33rd international conference on Very large data bases*, p. 1128–1137.

Castellanos, M., De Medeiros, A. K. A., Mendling, J., Weber, B., Weijters, A. J. M. M., (2009), "Business Process Intelligence", *Handbook of Research on Business Process Modeling*, Information Science Reference.

Chaudhuri, S., Dayal, U., (1997), "An overview of data warehousing and OLAP technology", *SIGMOD Rec.*, v. 26, n. 1 (mar.), p. 65–74.

Chun-Qin Gu, Hui-You Chang, Yang Yi, (2008), "Workflow mining: Extending the αalgorithm to mine duplicate tasks". In: *International Conference on Machine Learning and CyberneticsInternational Conference on Machine Learning and Cybernetics*, p. 361-368

Correia, J., (2002), *BAM: A Composite Market Changing the Way Enterprises Work*, Gartner. Disponível em: http://www.gartner.com/id=354280.

da Cruz, J. I. B., Ruiz, D., (2008), "Uma experiência em mineração de processos de manutenção de software". In: *Companion Proceedings of the XIV Brazilian Symposium on Multimedia and the Web*, p. 247–253, New York, NY.

Dongen, B. F. van, van der Aalst, W. M. P., (2005), "A Meta Model for Process Mining Data". *EMOI-INTEROP*

Van Dongen, B. F., De Medeiros, A. K. A., Verbeek, H. M. W., Weijters, A. J. M. M., van der Aalst, W. M. P., (2005), "The ProM framework: A new era in process mining tool support". In: *Lecture Notes in Computer ScienceApplications and Theory of Petri Nets 2005*, p. 444-454, Miami, USA.

Dumas, M., Aalst, W. M. van der, Hofstede, A. H. ter, (2005), *Process Aware Information Systems: Bridging People and Software Through Process Technology*. 1 ed. Wiley-Interscience.

Earls, A., (2011). BPMN 2.0: The emerging star of business process modeling. Disponível em: http://www.ebizq.net/topics/bpm_process_modeling/features/13202.html. Acesso em: 8 fev 2012.

Elmagarmid, A. K., Ipeirotis, P. G., Verykios, V. S., (2007), "Duplicate Record Detection: A Survey", *IEEE Trans. on Knowl. and Data Eng.*, v. 19, n. 1 (jan.), p. 1–16.

Esposito, P. M., Vaz, M. A. A., Rodrigues, S. A., Souza, J. M. de, (2012), "MANA: Identifying and Mining Unstructured Business Processes". *Aceito para publicação em BPI 2012 - 8th International Workshop on Business Process Intelligence, a ser relizado em conjunto ao BPM 2012 - 10th International Conference on Business Process Management*, Tallinn, Estônia.

Esposito, P. M., Vaz, M. A. A., Souza, J. M. de, Terres, L., (2011), "Uma abordagem para a mineração dos processos de uma universidade". In: *Anais do SBSI 2011 - VII Simpósio Brasileiro de Sistemas de Informação.WBPM 2011 - V Workshop de Gestão de Processos de Negócio*, p. 485-492, Salvador.

Fluxicon, (2012). Nitro. Disponível em: http://www.fluxicon.com/nitro/. Acesso em: 3 abr 2012.

Forrester, (2010), *The Forrester Wave: Enterprise Business Intelligence Platforms, Q4 2010*, Forrester.

Gansner, E. R., Koutsofios, E., North, S. C., Vo, K.-P., (1993), "A Technique for Drawing Directed Graphs", *IEEE Trans. Softw. Eng.*, v. 19, n. 3 (mar.), p. 214–230.

Gartner, (2010), *Magic Quadrant for Business Process Analysis Tools*, Gartner.

Gottschalk, F., Aalst, W. M. P., Jansen-Vullers, M. H., (2008), "Merging Event-Driven Process Chains", In: Meersman, R., Tari, Z. [orgs.] (eds), *On the Move to Meaningful Internet Systems: OTM 2008*, , chapter 5331, Berlin, Heidelberg: Springer Berlin Heidelberg, p. 418-426.

Graphviz, (2011). Graphviz - Graph Visualization Software. Disponível em: http://www.graphviz.org/. Acesso em: 20 mar 2012.

Greco, G., Guzzo, A., Pontieri, L., Sacca, D., (2006), "Discovering Expressive Process Models by Clustering Log Traces", *IEEE Transactions on Knowledge and Data Engineering*, v. 18 (ago.), p. 1010–1027.

Grigori, D., Casati, F., Castellanos, M., Dayal, U., Sayal, M., Shan, M.-C., (2004), "Business Process Intelligence", *Computers in Industry*, v. 53, n. 3 (abr.), p. 321-343.

Gunther, C. W., (2009), *XES Standard Definition*, Fluxicon Process Laboratories. Disponível em: www.xes-standard.org.

Günther, C. W., (2009), *OpenXes - Developer Guide*, Fluxicon Process Laboratories. Disponível em: http://www.xes-standard.org/openxes/start.

Günther, C. W., van der Aalst, W. M. P., (2006), "A generic import framework for process event logs". In: *Proceedings of the 2006 international conference on Business Process Management Workshops*, p. 81–92, Berlin, Heidelberg.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H., (2009), "The WEKA data mining software: an update", *ACM SIGKDD Explorations Newsletter*, v. 11 (nov.), p. 10–18.

Hammer, M., Champy, J., (1994), *Reengineering the Corporation: A Manifesto for Business Revolution*. Reprint ed. Harperbusiness.

Hearst, M., (2006), "Design Recommendations for Hierarchical Faceted Search Interfaces". *ACM SIGIR Workshop on Faceted Search*, Seattle, USA.

IDS Scheer AG, (2008), *Aris Process Performance Manager - Process Analysis User Guide*, IDS Scheer AG.

IIBA, (2009), *A Guide to the Business Analysis Body of Knowledge*®. 2nd ed. International Institute of Business Analysis.

Inmon, W. H., (2005), *Building the Data Warehouse*. 4 ed. Wiley.

Kobielus, J., (2010). What's Not BI? Oh, Don't Get Me Started....Oops Too Late...Here Goes.... | Forrester Blogs. Disponível em: http://blogs.forrester.com/james_kobielus/10-04-30-what%E2%80%99s_not_bi_oh_don%E2%80%99t_get_me_startedoops_too_latehere_goes. Acesso em: 12 fev 2012.

Lawrence, P., (1997), *Workflow Handbook 1997*. 1 ed. Wiley.

Li, Liu, D., Yang, B., (2007), "Process mining: Extending -algorithm to mine duplicate tasks in process logs", *Advances in Web and Network Technologies and Information Management*, v. 4537, p. 396-407.

Marchionini, G., (2006), "Exploratory search: from finding to understanding", *Commun. ACM*, v. 49, n. 4 (abr.), p. 41–46.

McCoy, D., (2002), *Business Activity Monitoring: Calm Before the Storm*, Gartner.

De Medeiros, A. K. A., Van Dongen, B. F., Van Der Aalst, W. M. P., Weijters, A. J. M. M., (2004), "Process Mining: Extending the α-algorithm to Mine Short Loops", *Eindhoven University of Technology, Eindhoven.*

De Medeiros, A. K. A., Guzzo, A., Greco, G., Van Der Aalst, W. M. P., Weijters, A. J. M. M., Van Dongen, B. F., Saccà, D., (2007), "Process mining based on clustering: a quest for precision". In: *Proceedings of the 2007 international conference on Business process management*, p. 17–29, Berlin, Heidelberg.

Medeiros, A. K. A., Weijters, A. J. M. M., der Aalst, W. M. P., (2006), "Genetic Process Mining: A Basic Approach and Its Challenges", In: Bussler, C. J., Haller, A. [orgs.] (eds), *Business Process Management Workshops*, , chapter 3812, Berlin, Heidelberg: Springer Berlin Heidelberg, p. 203-215.

Microsoft, (2010). Microsoft SQL Server. Disponível em: http://www.microsoft.com/sqlserver/pt/br/default.aspx. Acesso em: 20 mar 2012.

Ministério do Planejamento, Orçamento e Gestão, (2004), *Controle de Processos e Documentos - Manual do Sistema*, Ministério do Planejamento, Orçamento e Gestão.

Morris, T., (2012). Google Refine - Clustering in Depth. Disponível em: http://code.google.com/p/google-refine/wiki/ClusteringInDepth. Acesso em: 10 jun 2012.

Nesamoney, D., (2004). BAM: Event-Driven Business Intelligence for the Real-Time Enterprise. *Information Management Magazine*. Disponível em: http://www.information-management.com/issues/20040301/8177-1.html. Acesso em: 18 fev 2012.

OKTLAB, (2011). OKT Process Mining Suite. Disponível em: http://oktlab.openknowtech.it/OKTLAB/en/newsview.wp?contentId=NEW444. Acesso em: 29 jul 2012.

OMG, (2011), *Business Process Model and Notation (BPMN), Version 2.0*, OMG.

OMG, (2012). Introduction to OMG UML. Disponível em: http://www.omg.org/gettingstarted/what_is_uml.htm. Acesso em: 8 fev 2012.

Oracle, (2009). JavaServer Faces Technology. Disponível em: http://www.oracle.com/technetwork/java/javaee/javaserverfaces-139869.html. Acesso em: 30 mar 2012.

Perceptive Software, (2011), *Perceptive Reflect - Product Datasheet* Disponível em: http://www.perceptivesoftware.com/images/PerceptiveSoftware_Product_PerceptiveReflect.pdf.

Rodrigues, T., (2011), *WHYSEARCH: Um mecanismo causal e temporal de encadernamento de notícias*. Dissertação de Mestrado, UFRJ.

Rós, E., Baldam, R., Co, F., Lorenzoni, L., (2009), "Os ciclos de BPM (Gerenciamento de Processos de Negócios): Uma proposta de ação integrada". *ADM 2009 - Congresso Internacional de Administração*, Ponta Grossa.

Rozinat, A., (2011). How Process Mining Compares to BI — Flux Capacitor. Disponível em: http://fluxicon.com/blog/2011/01/how-pm-compares-to-bi/. Acesso em: 12 fev 2012.

Schedlbauer, M., (2010), *The Art of Business Process Modeling: The Business Analyst's Guide to Process Modeling with UML & BPMN*. CreateSpace.

Sharon, L. T., Bitzer, M., Kamel, M. N., (1997), "Workflow reengineering: a methodology for business process reengineering using workflow management technology". In: *Proceedings of the Thirtieth Hawaii International Conference on System Sciences, 1997Proceedings of the Thirtieth Hawaii International Conference on System Sciences, 1997*, p. 415-426 vol.4, Hawaii, USA.

Song, M. S., Günther, C. W., van der Aalst, W. M. P., (2008), "Trace Clustering in process mining". In: *Business Process Management Workshops*, p. 109-120, Milano, Italy.

Szathmary, L., (2011). GraphViz Java API. Disponível em: http://www.loria.fr/~szathmar/off/projects/java/GraphVizAPI/index.php. Acesso em: 20 mar 2012.

Valle, R., de Oliveira, S. B., (2009), *Análise e Modelagem de Processos de Negócio: Foco na Notação BPMN*. Editora Atlas.

Vassiliadis, P., (2009), "A Survey of Extract–Transform–Load Technology", *International Journal of Data Warehousing and Mining*, v. 5, n. 3 (33.), p. 1-27.

Veiga, G. M., Ferreira, D. R., (2010), "Understanding Spaghetti Models with Sequence Clustering for ProM". In: *Business Process Management WorkshopsBPM 2009 International Workshops*, p. 92–103, Ulm, Germany.

webMethods, (2006), *Business Activity Monitoring (BAM) - The New Face of BPM*, webMethods.

Weijters, A. J. M. M., van der Aalst, W. M. P., De Medeiros, A. K. A., (2006), "Process Mining with the HeuristicsMiner Algorithm", *BETA Working Paper Series, WP 166, Eindhoven University of Technology, Eindhoven.*

Wen, L., Wang, J., Sun, J., (2006), "Detecting Implicit Dependencies Between Tasks from Event Logs", *Frontiers of WWW Research and Development - APWeb 2006*, , p. 591-603.

Weske, M., (2007), *Business Process Management: Concepts, Languages, Architectures*. 1 ed. Springer.

White, R. W., Kules, B., Drucker, S. M., schraefel, m. c., (2006), "Supporting exploratory search: Introduction", *Commun. ACM*, v. 49, n. 4 (abr.), p. 36–39.

White, S., (2005), *Introduction to BPMN*, IBM.

Yee, K.-P., Swearingen, K., Li, K., Hearst, M., (2003), "Faceted metadata for image search and browsing". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, p. 401–408, New York, NY, USA.