Similaridade para Avaliação de Riscos em Planos de Mudança de TI

Luis Armando Bianchin[1], Juliano Araujo Wickboldt[1], Ricardo Luis dos Santos[1], Roben Castagna Lunardi[1], Bruno Lopes Dalmazo[1], Fabricio Girardi Andreis[1], Weverton Luis da Costa Cordeiro[1], Abraham Lincoln Rabelo de Sousa[1], Lisandro Zambenedetti Granville[1], Luciano Paschoal Gaspary[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre - RS
{labianchin, jwickboldt, rlsantos, rclunardi, bldalmazo, fgandreis, wlccordeiro, rabelo, granville, paschoal}@inf.ufrgs.br

Abstract. *The proper management of IT infrastructures is essential for organizations that aim to deliver high quality services. Given the dynamics of these environments, changes become imminent. In some cases these changes might raise failures that may cause disruption to services affecting the business continuity, which makes necessary the evaluation of the risks associated with changes before their actual execution. Taking advantage of information from past deployed changes it's possible to estimate the risks for recently planned ones. Thereby, in this paper, we propose a solution to weigh the information available from past executed changes by the similarity calculated in relation with the analyzed change. A prototype system was developed in order to evaluate the efficacy of the solution in an emulated IT infrastructure. The results show that the solution is capable of capturing similarity among changes, improving the accuracy of risk assessment for IT change planning.*

Resumo. *O gerenciamento apropriado de infra-estruturas de TI é fundamental para organizações que buscam oferecer serviços de alta qualidade. Dada a dinâmica desses ambientes, mudanças tornam-se iminentes. Em alguns casos as mudanças causam falhas que podem afetar a disponibilidade dos serviços afetanto a continuidade do negócio, o que torna necessário avaliar os riscos associados a essas mudanças antes que sejam executadas. Utilizando-se informações de mudanças implantadas anteriormente é possível estimar os riscos de mudanças recém planejadas. Para isso, neste trabalho, é proposta uma solução para ponderar os dados disponíveis sobre mudanças passadas pela similaridade que possuem em relação à mudança analisada. Um protótipo foi desenvolvido a fim de avaliar a eficácia da solução numa infra-estrutura de TI emulada. Os resultados mostram que a solução é capaz de capturar similaridade entre diferentes mudanças, melhorando a precisão das estimativas de risco no planejamento de mudanças.*

1. Introdução

Organizações que buscam oferecer serviços de alta qualidade normalmente precisam tratar o aumento de tamanho e complexidade de suas infra-estruturas de tecnologia da informação (TI). Infra-estruturas modernas incluem Itens de Configuração (ICs) que variam

de elementos físicos como servidores, estações de trabalho, dispositivos móveis e roteadores, a elementos lógicos como pacotes de *software* e serviços de rede. A fim de auxiliar tais organizações a empregar um gerenciamento racional de suas infra-estruturas de TI, o *Office of Government Commerce* (OGC) introduziu o *Information Technology Library* (ITIL) [ITIL 2009]. O ITIL apresenta essencialmente de um conjunto de boas práticas e processos cujo objetivo é guiar o gerenciamento apropriado de recursos e serviços de TI.

O *gerenciamento de mudanças* é um dos principais tópicos abordados pelo ITIL e que define como mudanças devem ser conduzidas numa infra-estrutura de TI. O ITIL define que mudanças devem ser especificadas de forma declarativa, em documentos chamados *Requests for Change* (RFCs). Tais RFCs devem então ser processadas, manual ou automaticamente, a fim de gerar Planos de Mudança (PM), os quais são *workflows* de ações que, quando executados, levarão a infra-estrutura de TI gerenciada a um novo estado funcional que será consistente com as mudanças originalmente expressas na RFC. Porém, devido a problemas imprevisíveis que podem ocorrer durante o desenrolar das mudanças, os quais podem causar interrupções nos serviços da infra-estrutura de TI, é conveniente avaliar os riscos associados aos PM antes de sua execução sobre a infra-estrutura gerenciada.

A avaliação de risco em gerenciamento de mudanças de TI é uma área de pesquisa recente que apresenta desafios bastante interessantes. Um deles, sendo de especial interesse na pesquisa apresentada neste artigo, reside no fato de que metodologias de estimativa de riscos em mudanças utilizando abordagem baseadas na análise dos históricos de execuções passadas de PMs requerem a execução recorrente de um mesmo PM para que se possa extrair resultados relevantes. No caso de PMs recém especificados, ou seja, sem histórico de execuções para avaliação, tal cômputo não seria viável. Isso conduz a situação em que PMs definidos para mudanças nunca executadas anteriormente não podem ter seu potencial de afetar os recursos de TI observados; operadores de TI não têm alternativa exceto executar os novos PMs e lidar de forma reativa com os problemas que podem vir a ocorrer durante a execução.

Neste trabalho, porém, argumentamos que a avaliação de risco pode ainda ser feita se os riscos de novos PMs forem computados considerando execuções passadas de PMs *similares*. Assim, neste trabalho é investigada uma solução para medir a similaridade entre atividades de PMs. Nossa abordagem consiste em comparar as atividades de um novo PM de interesse com as atividades de PMs já empregados anteriormente na infra-estrutura de TI gerenciada, e assim selecionar atividades similares através do uso de um algoritmo específico para este fim. Em seguida, execuções passadas dos PMs existentes são observadas, ponderadando-as pelas similaridades encontradas, e assim permitindo uma estimativa com maior precisão da probabilidade de falha das atividades do novo PM. A solução foi avaliada em um estudo de caso conduzido sobre uma infra-estrutura de TI emulada, a fim de avaliar seu pontecial em capturar atividades similares.

O restante deste artigo está organizado da seguinte forma. Na Seção 2 são apresentados os trabalhos relacionados ao tema desta pesquisa. Na Seção 3 alguns conceitos e definições usados na solução são explicados. A solução proposta é detalhada na Seção 4. Na Seção 5 um estudo de caso é desenvolvido usando a abordagem introduzida na seção anterior. Finalmente, na Seção 6 o artigo é concluído com considerações finais e indicações de trabalhos futuros.

2. Trabalhos Relacionados

Gerenciamento de risco é um tópico que tem sido amplamente discutido em áreas tão diversas quanto engenharia, medicina e economia. Risco é um conceito relacionado com o potencial de eventos incertos ocorrerem, normalmente com efeitos negativos, que afetam a realização dos objetivos dos negócios [Office of Government Commerce 2007]. Especialmente em gerenciamento de mudanças, risco é um aspecto importante que deve ser analisado, já que mudanças mal implementadas podem resultar em falhas que causam interrupções em serviços críticos para a continuidade dos negócios. Para promover a análise de riscos em gerenciamento de mudanças, as boas práticas do ITIL [ITIL 2007] sugerem que riscos devem ser avaliados e mitigados antes de uma mudança ser aprovada, reduzindo assim tanto a chance de ocorrer eventos negativos como também minimizando o impacto que esses eventos podem ter sob a infra-estrutura gerenciada.

Setzer *et al.* [Setzer *et al.* 2008] e Sauvé *et al.* [Sauvé *et al.* 2007] pesquisaram sobre a análise de risco no processo de planejamento do agendamento da execução de RFCs. Guiados por objetivos de negócios, a abordagem dos autores baseia-se na determinação de prioridades de execução de RFCs potencialmente concorrentes, com o objetivo de minimizar os riscos e os custos de implantação sobre os serviços das empresas. De acordo com os autores, o tempo elevado de indisponibilidade nos serviços durante a implantação de mudanças pode prejudicar severamente os serviços de negócio. Assim, são analisadas estratégia de implantação de RFCs considerando o impacto que cada RFC do conjunto pode ter sobre o negócio.

Em outro trabalho, Wickboldt *et al.* [Wickboldt *et al.* 2009b], a fim de permitir eventuais ajustes em uma RFC antes de sua aprovação, propuseram uma solução para avaliar os riscos já na fase de planejamento de mudanças, considerando tanto a probabilidade de ocorrerem falhas quanto a relevância dos elementos da infra-estrutura de TI envolvidos, o que permite compreender também o impacto de eventuais falhas. Para a estimativa de probabilidade de falhas, os autores usaram registros de execuções passadas como mecanismo para encontrar PMs suficientemente parecidos com o PM que estava em análise, levando-se em conta informações como a quantidade de falhas e execuções, além da similaridade dos planos envolvidos. Neste trabalho, porém, a busca por PMs similares foi realizada de forma extremamente rudimentar, sem considerar aspectos importantes que permitiram identificar similaridades de forma mais adequada.

Na tentativa de se determinar mais precisamente a similaridade entre *workflows* de mudanças, estes poderiam ser modelados como grafos dirigidos e então terem sua similaridade computada a partir de técnicas já utilizadas em grafos, como por exemplo no trabalho de Chartrand *et al.* [Chartrand *et al.* 1998]. Tais técnicas visam atingir o isomorfismo entre os grafos a partir da verificação da quantidade de operações necessárias, sobre e arcos e nodos, para transformar um grafo em outro. Porém, além de serem mais complexas, essas técnicas buscam uma comparação considerando apenas nodos e arcos, não levando em conta aspectos semânticos fundamentais em *workflows* de mudança, como a seqüencialidade e paralelismo entre atividades.

Outros autores investigaram a similaridade entre *workflows* considerando noções de equivalência de traço - como na pesquisa de Hidders *et al.* [Hidders *et al.* 2005] - e bissimulação - como proposto por Van der Aalst *et al.* [Van der Aalst e Basten 2002]. Porém, tais equivalências não são aplicáveis ao contexto de nossa pesquisa porque ofere-

cem uma resposta com granularidade muito baixa; ou *workflows* são equivalentes ou não são. Outras pesquisas, como as de Van Dongen *et al.* [Van Dongen *et al.* 2008] e Van der Aalst *et al.* [Van der Aalst *et al.* 2006], investigaram a similaridade de *workflows* de processos considerando os logs de execução para comparar o comportamento dos *workflows* analisados. Tal abordagem, porém, não se aplica ao contexto de avaliação de risco porque ainda não se sabe o comportamento do novo PM que está sendo analisado ao ser executado; na realidade pretende-se prever qual será o seu comportamento quando o PM for executado.

Por fim, Wombacher e Rozie [Wombacher e Rozie 2006] compararam vários métodos de similaridade aplicáveis a autômatos e grafos, avaliando seu uso em *workflows*. Seguindo essa linha de pesquisa, Li *et al.* [Li *et al.* 2008] propuseram uma medida de similaridade de modelos de processos em que usa-se técnicas de lógica digital para calcular esse escore. Porém, mesmo que PMs sejam compostos de *workflows*, é importante analisar o detalhamento das atividades que compõem estes *workflows*, dando importância também aos participantes envolvidos nas atividades. Desse modo, faz-se necessário construir uma solução para cálculo de similaridade que leve em conta também outros aspectos que favoreçam o contexto de análise de risco.

3. Definições

A fim de fornecer embasamento teórico à solução proposta neste trabalho, inicialmente, nesta seção, são revisados e formalizados alguns conceitos importantes propostos em trabalhos anteriores. Além disso, são introduzidos alguns novos conceitos utilizados na solução que será apresentada na seção seguinte.

3.1. Atividade

Uma atividade descreve uma única operação envolvendo elementos de *software*, *hardware* e demais Itens de Configuração (ICs), que pode ser realizada de forma automatizada ou manual - nesse caso envolvendo humanos - e cujo objetivo é modificar os ICs de forma a contemplar as mudanças descritas em uma RFC. As atividades são organizadas nos PMs na forma de um *workflow* que determina: a ordem de execução das atividades, restrições temporais entre elas e possíveis paralelismos. As operações executadas pelas atividades afetam os seus participantes, por exemplo, ao se instalar ou remover pacotes de *software* em computadores, ao se alterar as configurações de roteadores ou ao se editar as regras de *firewall*. No caso de atividades manuais, recursos humanos também são associados às atividades na forma de participantes.

Neste trabalho, uma atividade é formalizada como uma tupla: $A = \langle \Omega, \lambda \rangle$, onde:

- Ω é a operação realizada pela atividade (*e.g.*, instalação, atualização, desinstalação e configuração);
- λ é o conjunto de elementos participantes da atividade (*e.g.*, humanos, elementos de *hardware*, *software* e demais ICs).

3.2. Tipos de Falha

Para todos os PMs implantados sobre a infra-estrutura de TI, é armazenado o registro (*log*) das atividades executadas. Esses registros contêm os traços de execução dos PM, permitindo a posterior recuperação das informações do *workflow* e a ordem em que as

atividades foram executadas. Além disso, é possível extrair dos *logs* informações sobre o êxito ou fracasso das execuções e, quando há falhas, estas são classificadas em seis categorias ou Tipos de Falha (TF) [Wickboldt *et al.* 2009c]: (*i*) Falha de Atividade (FA), (*ii*) Falha de Recurso (FR), (*iii*) Falha de Humano (FH), (*iv*) Falha de Tempo (FT), (*v*) Intervenção Externa (IE) e (*vi*) Violação de Restrição (VR).

Um aspecto importante ao se classificar falhas ocorridas em mudanças é que dessa forma se permite associar uma falha ao IC que a ocasionou. Por exemplo, considerando o caso de uma atividade de instalação de *software* sobre um determinado computador; se a falha ocorrida é uma FA, diz-se que o elemento que a provocou foi o *software*, enquanto que se a falha é classificada como FR (Falha de Recurso), esse evento ficará associado ao *hardware*. Como o foco deste trabalho está no cálculo da *similaridade* de PMs, e também por medida de simplificação, serão consideradas apenas Falha de Atividade (FA), que são ocasionadas por problemas inerentes às atividades do PM, tais como exceções geradas durante a instalação ou configuração de um software. Informações sobre os outros cinco Tipos de Falha (TF) são descritas no trabalho de Wickboldt *et al.* [Wickboldt *et al.* 2009c].

Neste trabalho, considera-se que as mudanças realizadas sobre a infra-estrutura de TI são controladas e documentadas conforme as recomendações feitas pelo ITIL. Em trabalhos passados deste grupo de pesquisa, uma solução fim-a-fim de sistema de planejamento e execução de mudanças foi proposta [Cordeiro *et al.* 2008, Wickboldt *et al.* 2009a]. Com auxilio de um sistema como esse, é possível manter de forma organizada o histórico das mudanças realizadas sobre cada CI. A detecção e tratamento de falhas ocorridas durante as mudanças estão fora do escopo deste trabalho, porém é importante que esses eventos sejam devidamente documentados, seja este um processo automatizado ou executado manualmente durante a revisão das mudanças, a fim de permitir futuras estimativas de riscos.

3.3. Workflow Influencial

Chamamos Workflow Influencial o subconjunto de atividades de um *workflow* que podem influenciar a execução de uma dada atividade dentro de um mesmo PM. Considera-ses neste trabalho que uma atividade A pode influenciar a execução (eventualmente também as falhas) de outra atividade B quando: (*i*) A antecede B no *workflow*, ou seja, para que B possa ser executada A deve ter sido concluída primeiro, ou (*ii*) A está em paralelo com B, sendo assim a ordem suas execuções não é determinística.

Intuitivamente, no que diz respeito à análise de riscos, o primeiro caso captura situações em que a falha de uma atividade B foi causada indiretamente por problemas ocorridos em atividades que a antecederam, enquanto que o segundo caso captura problemas ocasionados por execuções em paralelo onde pode haver, por exemplo, disputa por recursos compartilhados. Em outras palavras, o Workflow Influencial é o próprio *workflow* excetuando as atividades que ocorrem após uma dada atividade, já que a execução dessa atividade não pode receber interferência das atividades que vêm a seguir. Além disso, a atividade objeto da análise também é incluída no seu Workflow Influencial juntamente com as atividades que a antecedem ou estão em paralelo com ela. Ademais, as transições entre as atividades do *workflow* original são preservadas no Workflow Influencial.

3.4. Similaridade

Uma métrica de similaridade objetiva estimar quão parecidas duas entidades de qualquer natureza são. Basicamente, existem duas propriedades relacionadas com as medidas de similaridade que são muito interessantes para a solução descrita na próxima seção:

- **Comutatividade:** determina que o valor de similaridade entre X e Y é igual ao valor de similaridade de Y e X;
- **Intervalo de 0 a 1:** o escore de similaridade varia de 0, totalmente diferentes, a 1, exatamente iguais.

Métricas de cálculo de similaridade que comumente respeitam essas propriedades são as utilizadas para comparação de *strings*, as quais são largamente empregadas para encontrar semelhanças entre textos, análises de DNA, mineração de dados, entre outros fins. Um conceito muito utilizado no cálculo de similaridade de *strings* que vem a ser bastante importante para a solução proposta neste trabalho é o conceito de distância. Basicamente, a distância entre duas entidades representa o número de operações básicas para transformar uma entidade noutra. De fato, existem diferentes medidas de distância de *strings*. Em particular, cabe ressaltar a distância de edição (ou Distância de Levenshtein) entre duas palavras que representa a menor quantidade de inserções, substituições e supressões de símbolos para transformar uma palavra em outra [Levenshtein 1966]. Usando um valor de distância d, a similaridade pode ser obtida subtraindo-se este valor a partir da diferença máxima entre as duas entidades m e dividindo por essa diferença máxima, ou seja, $sim = \frac{m-d}{m}$ [Wombacher e Rozie 2006].

Já para cálculo de coeficiente de similaridade entre conjuntos, outras métricas também foram estudas [Cohen *et al.* 2003], como Jaccard, MongeElkan [Monge e Elkan 1996] e SoftTFIDF. Em especial, convém detalhar o funcionamento do índice de Jaccard: a partir de dois conjuntos C_1 e C_2, pode-se calcular o escore de similaridade pela relação entre o número de elementos comuns e a quantidade total de elementos em ambos os conjuntos.

$$sim\_jaccard = \frac{|C_1 \cap C_2|}{|C_1 \cup C_2|} \quad (1)$$

O índice de Jaccard (Equação 1), bem como os demais conceitos e definições descritos nesta seção, servem de base para a solução de cálculo de similaridade em planos de mudança apresentada no seguimento deste artigo.

4. Solução Proposta

A solução para cálculo de similaridade proposta neste artigo é capaz de analisar um *workflow*, atividade por atividade, encontrando dentro de uma base de dados de *workflows* (*e.g.*, uma base de dados de PMs previamente executados) atividades similares às analisadas considerando dois aspectos: (*i*) a similaridade entre as duas atividades e seus participantes, chamada neste artigo de Similaridade Pontual e (*ii*) a similaridade dos Workflows Influenciais de ambas as atividades a fim de capturar a similaridade do contexto ou ambiente em que foram executadas, chamada de Similaridade de Workflows.

A Figura 1 apresenta o fluxo de informações utilizado pela solução desde a leitura do *workflow* a ser analisado, a seleção das atividades similares a partir dos Registros de Execução, o processamento dessas informações, até a composição de um relatório com os resultados da análise de similaridade das atividades. Os algoritmos que realizam os cálculos de similaridade são apresentados na continuação desta seção.

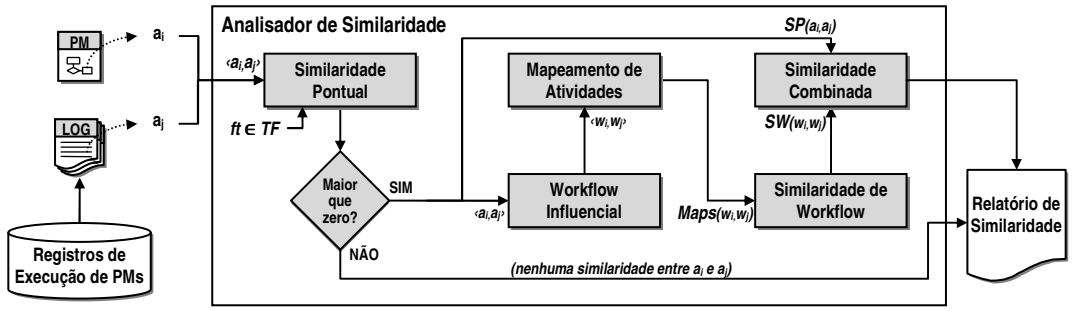

Figura 1. Fluxo de Informações da Solução

4.1. Similaridade Pontual

Usando a definição de atividade (Seção 3.1), pode-se estabelecer um algoritmo para comparar isoladamente duas atividades e determinar quão similares estas são no que diz respeito a operação realizada e participantes envolvidos. A esse algoritmo é atribuído o nome de Similaridade Pontual (SP).

Algoritmo 1 SIMILARIDADEPONTUAL

Entradas: Atividades $a_1 = \langle \Omega_1, \lambda_1 \rangle$ e $a_2 = \langle \Omega_2, \lambda_2 \rangle$, e um tipo de falha $ft \in TF$
Saídas: Similaridade Pontual das atividades a_1 e a_2 segundo o tipo de falha ft

1: **se** $\Omega_1 \neq \Omega_2$ **então**
2: $\quad SP \leftarrow 0$
3: **senão se** elementos do tipo ft de a_1 e a_2 são diferentes **então**
4: $\quad SP \leftarrow 0$
5: **senão**
6: $\quad SP \leftarrow \frac{|\lambda_1 \cap \lambda_2|}{|\lambda_1 \cup \lambda_2|}$
7: **fim se**
8: **retorna** SP

Obtém-se esse escore de SP a partir do Algoritmo 1, o qual se baseia fortemente no conceito de similaridade de Jaccard. Basicamente, para que duas atividades sejam significativamente relacionadas, elas devem executar a mesma operação (Linha 1) e possuir os mesmos participantes relacionados ao tipo de falha analisado (Linha 3), caso contrário, podemos afirmar que a SP entre essas atividades é zero, ou seja, são completamente diferentes. Caso contrário, o escore de similaridade será a razão entre a quantidade de participantes em comum e o total de participantes envolvidos em ambas as atividades, ou seja, Jaccard aplicado aos conjuntos de participantes de ambas as atividades (Linha 6). Assim, por exemplo, ao analisar a similaridade considerando Falhas de Atividade (FA), apenas execuções de atividades que realizam as mesmas operações sobre os mesmos elementos de *software* serão capturadas. É importante lembrar que neste trabalho são consideradas apenas FA, porém no algoritmo, ft pode assumir qualquer valor para um Tipo de Falha (descritos na Seção 3.2) pertencente a um conjunto TF. Sendo assim, para contemplar outros Tipos de Falha seria necessário repetir o mesmo algoritmo para cada um deles.

Na solução apresentada nesta seção, a SP determina o grau de semelhança entre as operações e os participantes de duas atividades. Sendo assim, na análise de riscos de Planos de Mudança a SP é utilizada para capturar as atividades relacionadas a partir

dos Registros de Execução. Isto é, caso a SP de duas atividades seja igual a zero, não é necessário executar nenhum dos outros algoritmos e as atividades são imediatamente consideradas diferentes. Assim, evita-se a criação dos Workflows Influenciais e a realização dos demais cálculos para todos PMs dos Registros de Execução, evitando desperdício de tempo e processamento.

4.2. Workflow Influencial e Mapeamentos de Atividades

Após calcular a SP do par de atividades sendo analisado (a_1 e a_2), caso esta resulte em um valor maior que zero, é necessário avançar na solução calculando os Workflows Influenciais (W_1 e W_2) das atividades e fazendo o Mapeamento de Atividades similares nos dois *workflows*. O algoritmo que monta o Workflow Influencial a partir de uma dada atividade não é detalhado neste artigo por limitação de espaço, porém seu funcionamento é bastante simples. Baseado na definição de Workflow Influencial (Seção 3.3), dada uma atividade a, basta extrair do *workflow* original a(s) atividade(s) que sigam as seguintes propriedades: (*i*) atividades que antecedem a, (*ii*) que estão em paralelo com a ou (*iii*) que seja a própria atividade a. As atividades que respeitam pelo menos uma das três propriedades são inseridas no Workflow Influencial na mesma ordem que aparecem no *workflow* original mantendo-se também as mesmas transições.

Uma vez calculados os Workflows Influenciais W_1 e W_2, procede-se então com o Mapeamento de Atividades similares nesses dois *workflows*. Tal mapeamento é necessário porque para o cálculo de similaridade de *workflows* é preciso comparar as atividades contidas em cada *workflow* par a par. Porém, não é possível determinar diretamente quais atividades são correspondentes nos *workflows* para formar os pares, uma vez que estas são objetos compostos de operação e participantes. Sendo assim o mapeamento é feito considerando as SP entre os pares de atividades.

Um Mapeamento de Atividades pode ser definido como $m = \langle a_1, a_2, sp \rangle$, no qual:

- a_1 é a atividade do primeiro *workflow*;
- a_2 é a atividade correspondente do segundo *workflow*;
- sp representa a Similaridade Pontual entre a_1 e a_2.

Sendo assim, pode-se criar um conjunto com todos os mapeamentos dos *workflows* W_1 e W_2 representado por $Maps(W_1, W_2)$ com a restrição de que as atividades pertencentes a esses mapeamentos não aparecem repetidamente. Em outras palavras, cada atividade de W_1 será mapeada em no máximo uma atividade de W_2, sendo que se não for encontrada uma atividade correspondente o mapeamento não é criado.

A partir disso, o Algoritmo 2 pode ser utilizado para capturar os mapeamentos entre as atividades de dois *workflows*. Conforme esse algoritmo, para obter os mapeamentos entre os Workflows Influenciais, pode-se construir duas pilhas: uma com as atividades de W_1 (Linha 2) e outra com as atividades de W_2 (Linha 7), em que a primeira atividade de cada *workflow* está na base da pilha. Ao retirar-se cada atividade da primeira pilha (Linha 11), retiram-se atividades da segunda a pilha (Linha 14) até que se tenha uma SP, entre estas, maior que zero. Assim, cada mapeamento é capturado e colocado num conjunto com todos os mapeamentos (Linha 21). Percebe-se que esse algoritmo possui uma complexidade de fator quadrático, já que num pior caso, todos os possíveis pares de atividades tem suas SPs calculadas.

Algoritmo 2 MAPS

Entradas: W_1, W_2 e um tipo de falha ft
Saídas: $Maps$ conjunto com mapeamentos $m = \langle a_1, a_2, sp \rangle$

1: $Maps \leftarrow \varnothing$
2: $P_1 \leftarrow novaPilha()$
3: **para cada** Atividade $a \in W_1$ */*partindo-se da atividade inicial*/* **faça**
4: $P_1.push(a)$
5: **fim para**
6: $P_2 \leftarrow novaPilha()$
7: **para cada** Atividade $a \in W_2$ */*partindo-se da atividade inicial*/* **faça**
8: $P_2.push(a)$
9: **fim para**
10: **enquanto** $P_1.naoVazio()$ **faça**
11: $a_1 \leftarrow P_1.pop()$
12: $L \leftarrow \varnothing$
13: **repita**
14: $a_2 \leftarrow P_2.pop()$
15: $sp \leftarrow SP(a_1, a_2, ft)$
16: **se** $sp = 0$ **então**
17: $L \leftarrow L \cup \{a_2\}$
18: **fim se**
19: **enquanto** $sp = 0$ **E** $P_2.naoVazio()$
20: **se** $sp > 0$ **então**
21: $Maps \leftarrow Maps \cup \{\langle a_1, a_2, sp \rangle\}$
22: **fim se**
23: **para cada** $a_2 \in L$ **faça**
24: $P_2.push(a_2)$
25: **fim para**
26: **fim enquanto**
27: **retorna** $Maps$

4.3. Similaridade de Workflows

Para obter a similaridade entre dois *workflows*, parte-se de definições de similaridade de outros autores que usam o conceito de distância para calcular a similaridade entre duas entidades [Li *et al.* 2008]. Essa distância é a soma de operações de inserção, remoção e substituição necessárias para transformar uma entidade noutra. Considerando o conceito de distância é possível derivar a seguinte equação para calcular a similaridade entre dois conjuntos A e B: $sim(A, B) = 1 - \frac{ins+rem+sub}{|A \cup B|}$. Sabendo que $ins + rem = |A \cup B| - |A \cap B|$, ou seja, que o número de inserções e remoções é igual ao número de atividades diferentes, pode-se derivar a seguinte equação: $sim(A, B) = \frac{|A \cap B| - sub}{|A \cup B|}$.

Seguindo esse raciocínio, pode-se usar conceito de Similaridade Pontual, realizando uma soma dos valores de sp dos mapeamentos ao invés do cardinal da intersecção dos conjuntos. Além disso, pode-se interpretar $|A_1 \cup A_2| = |A_1| + |A_2| - |A_1 \cap A_2|$, onde $|A_1 \cap A_2| = |Maps(W_1, W_2)|$, ou seja, pela soma do número de atividades contidas nos *workflows* subtraída pelo número de mapeamentos. Assim, para calcular o escore de simi-

laridade de *workflows* utiliza-se neste trabalho a Equação 2, onde A_1 e A_2 são conjuntos não ordenados que contém as atividades de W_1 e W_2 respectivamente.

$$SW_{(W_1,W_2)} = \frac{\sum_{m \in Maps} sp_m - subst(W_1, W_2)}{|A_1| + |A_2| - |Maps(W_1, W_2)|} \quad (2)$$

Para o cálculo do número de substituições necessário na Equação 2, utiliza-se a solução proposta por Li em [Li *et al.* 2008], em que a partir das atividades coincidentes em ambos *workflows* busca-se, através da lógica digital, encontrar o menor número de troca de posições (substituições) entre as atividades. Para isso, a partir do conjunto de atividades comum aos *workflows* montam-se matrizes com informações sobre a ordem de execução das atividades para cada *workflow*. A seguir, extrai-se uma expressão lógica com os conflitos encontrados a partir dessas duas matrizes. Com tamanho do maior termo, obtido a partir da minimização dessa expressão lógica, obtém-se o número de substituições.

Para calcular a distância de substituições para similaridade entre PMs, ao invés de utilizar as atividades comuns dos *workflows*, utilizam-se os mapeamentos obtidos, já que com estes é possível obter as atividades correspondentes nos *workflows*. Assim duas matrizes de ordenação são geradas, uma relativa a posição da primeira atividade dos mapeamentos e outra a partir da segunda. Após obter-se a expressão minimizada a partir dos conflitos das matrizes de ordenação dos mapeamentos, cada termo será valorado com a soma das Similaridades Pontuais dos mapeamentos que devem ser reposicionados.

4.4. Similaridade Combinada

Compondo os escores de Similaridade Pontual (SP) entre as atividades analisadas e de Similaridade de Workflows (SW) entre os Workflows Influenciais dessas atividades obtém-se o valor chamado neste trabalho de Similaridade de Combinada (SC), conforme a Equação 3.

$$SC_{(a_1,a_2)} = SP_{(a_1,a_2)} \cdot SW_{(infl(a_1),infl(a_2))} \quad (3)$$

Os valores obtidos para SC das atividades analisadas são organizados em um Relatório de Similaridade que é utilizado no cálculo de risco das atividades do Plano de Mudança analisado. As probabilidades de falha calculadas por uma solução de estimativa de riscos para as atividades extraídas dos Registros de Execução podem ser então ponderadas pelos valores de SC. Desse modo, mais peso é atribuído às probabilidades das atividades mais similares, permitindo a utilização adequada dos históricos de execução de PMs similares. Além disso, convém também salientar, que por tratar-se de um cálculo de escore de similaridade podem surgir falsos negativos que podem comprometer a análise de risco. Para amenizar esse problema, seria interessante estabelecer, por exemplo, um *threshold* de similaridade para inibir a inserção de ruído na análise causada por atividades com similaridade muito baixa.

5. Estudo de Caso

Para avaliação da eficácia da solução, propomos um estudo de caso, em que uma empresa de *hosting* oferece o serviço de instalação, configuração e hospedagem de servidores de

webmail utilizando a plataforma *Horde*. Seguindo as boas práticas, recomendadas pelo ITIL, RFCs para diferentes máquinas são especificadas conforme requistos de configuração especificados pelos clientes. Os PMs gerados a partir dessas RFCs estão representados na Figura 2. Todas essas RFCs possuem a instalação de um sistema operacional Linux - Fedora ou Debian -, a instalação e configuração de Apache e PHP, e a instalação e configuração do *Horde Webmail*. Além disso, algumas RFCs também fazem a instalação de banco de dados MySQL, outras de PostgreSQL, e ainda outras utilizam o banco de dados já instalado em alguma outra máquina, para fornecer o armazenamento ao serviço de *webmail*. Também, certas atividades variam de posição entre diferentes RFCs. Algumas dessas atividades são executadas manualmente (destacadas com hachuras na Figura 2), tais são a instalação do sistema operacional e configuração do *Horde Webmail*, sendo que um mesmo perfil de humano está associado a estas.

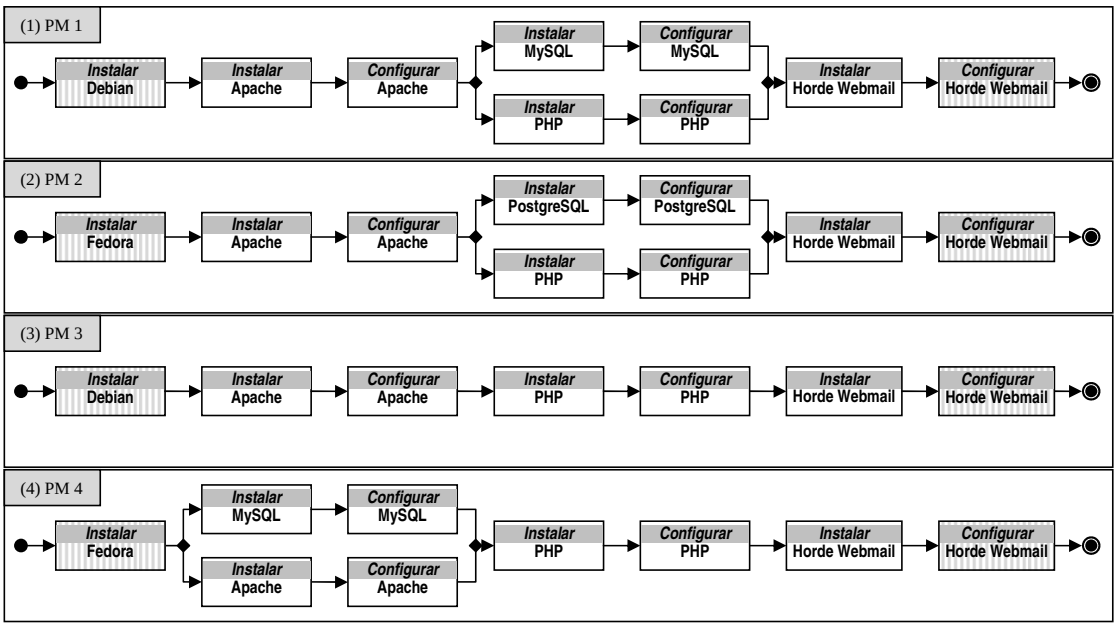

Figura 2. Planos de Mudança do Estudo de Caso

5.1. Análise das Similaridades

Observando-se a atividade de instalação da *Horde Webmail* do PM 4 e verificando a sua SC para Falhas de Atividade (FA) em relação as demais atividades desses PMs, nota-se que apenas as atividades *Instalar Horde Webmail* nos PMs têm SP maior que zero, já que são as únicas que possuem operação (*Instalar*) e participante de *software* (*Horde Webmail*) idênticos. Considerando essa atividade nos quatro PMs, a partir dos Workflows Influenciais obtidos com essas atividades - os quais são os próprios *workflows* excetuando-se a atividade *Configurar Horde Webmail* - obtém-se os mapeamentos apresentados na Tabela 1. Nessa tabela, estão apresentadas, por simplicidade, apenas a operação, participante de *software* e SPs em relação a essas atividades nos diferentes PMs.

Nesses mapeamentos, nota-se que as atividades *Instalar Fedora*, a qual possui um humano em comum com a mesma atividade do PM 4 apresenta SP de $0,50$, já que num total de quatro participantes apresenta dois em comum e dois distintos (*hardware*). Já as demais atividades, algumas não possuem mapeamentos, outras possuem SP de $0,33$, pois apenas o participante de *software* é idêntico, diferindo em 2 participantes de *hardware*.

Tabela 1. Mapeamentos das Atividades

	PM 1	PM 2	PM 3	PM 4
Instalar Fedora	-	0,50	-	1,00
Instalar Apache	0,33	0,33	0,33	1,00
Configurar Apache	0,33	0,33	0,33	1,00
Instalar MySQL	0,33	-	-	1,00
Configurar MySQL	0,33	-	-	1,00
Instalar PHP	0,33	0,33	0,33	1,00
Configurar PHP	0,33	0,33	0,33	1,00
Instalar Horde Webmail	0,33	0,33	0,33	1,00
Distância de Substituição	0,67	0,00	0,00	0,00
Soma das Similariades Pontuais	2,33	2,17	1,67	8,00
Tamanho Mapeamentos	7	6	5	8
Similaridade de Workflow	0,19	0,22	0,19	1,00
Similaridade Combinada	0,06	0,07	0,06	1,00

Além disso, na parte inferior da tabela, tem-se a soma dos mapeamentos e o número de mapeamentos, a serem usados para calcular a Similaridade de Workflow.

Através dos mapeamentos, pode ser feito o cálculo de distância de substituição, no qual verifica-se que apenas entre PMs 1 e 4 nas atividades *Instalar* e *Configurar MySQL* há conflitos de posição, entre estes obteve-se a distância de substituição no valor de $0,66$, ou seja, a soma das SP dessas atividades. A partir dos dados apresentados na tabela, podemos aplicar a Equação 2, de Similaridade de Workflow, obtendo os escores apresentados na mesma tabela. Para chegar ao escore final de SC, multiplicamos os escores de Similaridade de Workflow e Pontual da atividade analisada - *Instalar Horde Webmail* - seguindo a Equação 3.

5.2. Comparativo

Ao calcular-se as SCs de todas as atividades *Instalar Web Application* nos diferentes PMs, obtém-se os valores dispostos na mos a Tabela 2 (a). Para comparação, foram calculados os escores de similaridade para as mesmas atividades utilizando uma solução anterior proposta por Wickboldt *et. al* [Wickboldt *et al.* 2009b], cujos resultados encontram-se na Tabela 2 (b). Além disso, convém salientar que pelo fato de a similaridade ser comutativa, omite-se os valores da diagonal inferior da tabela.

Tabela 2. Matriz de Similaridades

(a)	(a) Similaridade Combinada				(b)	Similaridade anterior			
	PM 1	PM 2	PM 3	PM 4		PM 1	PM 2	PM 3	PM 4
PM 1	1,000	0,051	0,090	0,062	PM 1	1,000	0,313	0,396	0,438
PM 2		1,000	0,062	0,072	PM 2		1,000	0,313	0,396
PM 3			1,000	0,062	PM 3			1,000	0,313
PM 4				1,000	PM 4				1,000

Comparando-se os resultados obtidos com a métrica proposta por Wickboldt *et. al* aos resultados de Similaridade Combinada, percebe-se que o fato de esta nova métrica combinar os escores de Similaridade de Workflows e Pontual faz com que a similaridade

apresente valores menores em relação a solução anterior, uma vez que esta considerava apenas a similaridade da estrutura dos *workflows*. Isso é realmente interessante para a estimativa da probabilidade de falhas, pois quando se analisa riscos de PM que possuem seu próprio histórico de execução, as falhas desses planos terão muito mais peso do que as de outros planos apenas similares. Mesmo assim, os planos similares ainda possuem influência sobre o resultado final das probabilidades e, no caso de mudanças recém planejadas, serão determinantes para a obtenção dessa estimativa.Além disso, também verifica-se que pelo fato de a Similaridade Combinada levar em conta a posição das atividades no *workflow*, esta métrica apresenta um escore mais refinado em relação a solução anterior. Em certos casos isso pode alterar a ordem das atividades mais similares, como no caso em que na métrica anterior, PM 4 apresenta maior similaridade que PM 3 em relação a PM 1, já na nova métrica, ocorre a situação oposta.

6. Conclusão

Neste trabalho, foi proposta uma nova métrica para cômputo de similaridade de atividades de PMs visando aprimorar a análise de riscos baseada em dados históricos. Foi visto que nessa métrica de similaridade entre os principais aspectos levados em conta estão a importância de se considerar atividades que ocorrem antes da atividade analisada e a comparação dos participantes das atividades, relacionando-os aos seus tipos.

Os resultados obtidos a partir da execução de casos de testes numa infra-estrutura emulada, mostram a eficácia da solução em capturar os aspectos fundamentais de similaridade. Também a partir da comparação com solução anterior foi possível verificar que os escores obtidos apresentam uma maior amplitude, o qual reflete melhor as diferenças entre as atividades e torna a estimativa de probabilidade de falhas mais precisa.

Em trabalhos futuros, pretende-se ampliar os aspectos abordados, intrínsecos às características das infra-estruturas, refinando o conceito de Workflow Influencial. Além disso, pode-se estender a aplicabilidade da solução para outros contextos relacionadas como refinamento e alinhamento de planos, estimativa de custos e tempo, sempre aproveitando a base de informações históricas. A solução também pode ser usada numa abordagem evolutiva de infra-estruturas de TI, em que a partir de dados e decisões passadas, o sistema auxiliaria na tomada de decisões futuras.

O presente artigo foi alcançado em cooperação com a *Hewlett-Packard Brasil Ltda.* e com recursos provenientes da Lei de Informática (Lei nº 8.248, de 1991).

Referências

Chartrand, G., Kubicki, G., e Schultz, M. (1998). Graph similarity and distance in graphs. *Aequationes Mathematicae*, 55(1):129–145.

Cohen, W., Ravikumar, P., e Fienberg, S. (2003). A comparison of string distance metrics for name-matching tasks. Em *IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03)*, páginas 9–10.

Cordeiro, W. L. C., Machado, G. S., Daitx, F. F., *et al.* (2008). A Template-based Solution to Support Knowledge Reuse in IT Change Design. Em *11th IEEE/IFIP Network Operations and Management Symposium (NOMS 2008)*, páginas 355–362.

Hidders, J., Dumas, M., van der Aalst, W., ter Hofstede, A., e Verelst, J. (2005). When are two workflows the same? Em *Proceedings of the 2005 Australasian symposium on Theory of computing-Volume 41*, página 11. Australian Computer Society, Inc.

ITIL (2007). ITIL - Information Technology Infrastructure Library: Service Transition Version 3.0. Office of Government Commerce (OGC).

ITIL (2009). ITIL - Information Technology Infrastructure Library. Office of Government Commerce (OGC). Disponível em: http://www.itil-officialsite.com/. Acessado em: out. 2009.

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Em *Soviet Physics-Doklady*, volume 10.

Li, C., Reichert, M., e Wombacher, A. (2008). On measuring process model similarity based on high-level change operations. Em *27th International Conference on Conceptual Modeling (ER 2008)*, páginas 248–264.

Monge, A. e Elkan, C. (1996). The field matching problem: Algorithms and applications. Em *Second International Conference on Knowledge Discovery and Data Mining (KDD 96)*, páginas 267–270.

Office of Government Commerce (2007). Management of risk: Guidance for practitioners. Office of Government Commerce (OGC).

Sauvé, J., Santos, R. A., Almeida, R. R., Moura, A., *et al.* (2007). On the Risk Exposure and Priority Determination of Changes in IT Service Management. Em *18th IFIP/IEEE International Workshop on Distributed Systems: Operations and Management (DSOM 2007)*, páginas 147–158.

Setzer, T., Bhattacharya, K., e Ludwig, H. (2008). Decision support for service transition management Enforce change scheduling by performing change risk and business impact analysis. Em *11th IEEE/IFIP Network Operations and Management Symposium (NOMS 2008)*, páginas 200–207.

Van der Aalst, W. e Basten, T. (2002). Inheritance of workflows: an approach to tackling problems related to change. *Theoretical Computer Science*, 270(1-2):125–203.

Van der Aalst, W., de Medeiros, A., e Weijters, A. (2006). Process Equivalence: Comparing Two Process Models Based on Observed Behavior. Em *4th International Conference on Business Process Management (BPM 2006)*, volume 4102, páginas 129–144.

Van Dongen, B., Dijkman, R., e Mendling, J. (2008). Measuring similarity between business process models. Em *CAiSE*, páginas 450–464. Springer.

Wickboldt, J., Lunardi, R., Machado, G., *et al.* (2009a). Automatizando a Estimativa de Riscos em Sistemas de Gerenciamento de Mudanças em TI. Em *XXVII Simpósio Brasileiro de Redes de Computadores (SBRC 2009)*, páginas 423–436.

Wickboldt, J. A., Bianchin, L. A., Lunardi, R. C., *et al.* (2009b). Improving IT Change Management Processes with Automated Risk Assessment. Em *20th IFIP/IEEE International Workshop on Distributed System: Operations and Management (DSOM 2009)*, páginas 71–84.

Wickboldt, J. A., Machado, G. S., Cordeiro, W. L. C., *et al.* (2009c). A Solution to Support Risk Analysis on IT Change Management. Em *11th IFIP/IEEE International Symposim on Integrated Network Management (IM 2009)*, páginas 445–452.

Wombacher, A. e Rozie, M. (2006). Evaluation of workflow similarity measures in service discovery. *Service Oriented Electronic Commerce*, 7:26.