A Systematic Review about Activities Recommendation in Workflows

Adilson Lopes Khouri (Universidade de São Paulo, São Paulo, Brasil) – adilson.khouri.usp@gmail.com

Luciano Antonio Digiampietri (Universidade de São Paulo, São Paulo, Brasil) – luciano.digiampietri@gmail.com

Keywords: Workflows Construction, Activities Recommendation, Workflows Composition, Systematic Review

Abstract

Workflow management systems enable users without deep knowledge in computing connecting activities to build new software graphically. To harness the reuse in these systems, it is necessary knowing a large number of activities. In order to minimize this need, there are several proposals for workflows activities recommendation or automatic composition. This paper aims to identify the techniques used to recommend activities in workflows. For this, 20 papers were selected applying a systematic review, the methods found were analyzed to identify trends. It was found a tendency to use methods based on frequency, data provenance and information dependency. Another trend was the lack of formal and well-defined methodology to validate results.

Palavras-chave: Construção de *Workflows*, Recomendação de Atividades, Composição de *Workflows*, Revisão Sistemática

Resumo

Sistemas gerenciadores de *workflows* científicos permitem a usuários, sem grandes conhecimentos em computação, conectar atividades para construir novos *softwares* graficamente. Para possibilitar o reuso nesses sistemas é necessário conhecer um grande número de atividades. Para minimizar essa necessidade existem diversas propostas para recomendar atividades ou compor *workflows* automaticamente. O objetivo deste trabalho é identificar, através de uma revisão sistemática, técnicas existentes para recomendar atividades ou compor *workflows*. Para isto, 21 artigos selecionados através de uma revisão sistemática identificam os métodos encontrados. Foi encontrada uma tendência no uso de métodos baseados em frequência, proveniência e dependência de informação. Outra tendência encontrada foi a ausência de metodologias para testar e validar resultados.

Os autores agradecem a agência CAPES que contemplou o estudante com uma bolsa de mestrado possibilitando dedicação exclusiva e a elaboração deste artigo.

INTRODUÇÃO

Sistemas gerenciadores *de workflows* científicos permitem aos pesquisadores (sem grandes conhecimentos em computação) construir experimentos computacionais de forma simples, por exemplo, arrastando-se componentes gráficos (atividades) desenvolvidos anteriormente. Um requisito para aproveitar a capacidade de reutilização de componentes oferecida por estes sistemas é conhecer as atividades disponíveis, porém é inviável para qualquer usuário conhecer um número muito grande de atividades.

Atualmente há milhares de atividades disponíveis em repositórios como *myExperiment* que armazena mais de 2500 *workflows* (Roure, 2014) e *BioCatalogue* que disponibiliza 2464 serviços (Bhagat et al., 2014). O que tornou ainda mais necessário o desenvolvimento de mecanismos para auxiliar na construção de *workflows*. Dada essa inviabilidade, foram propostos diversos métodos para recomendar atividades ou compor automaticamente *workflows*, minimizando assim o requisito de se conhecer muitas atividades.

Existem duas abordagens principais utilizadas para auxiliar no processo de construção de *workflows*: (i) a composição semiautomática (recomendação), na qual sistemas de recomendação sugerem as atividades mais indicadas (segundo alguma métrica) para o *workflow* em construção (Bergmann & Gil, 2014); e (ii) a composição automática (composição) na qual dado um objetivo do usuário e/ou seus dados de entrada e saída desejada, o sistema tenta automaticamente criar um *workflow* que satisfaça a requisição do usuário (Ayadi & Lacroix, 2007).

Enquanto a recomendação de atividades é a abordagem indicada a usuários que conhecem o domínio da aplicação no qual o *workflow* será construído, a composição automática é indicada a usuários que apenas conhecem seus objetivos sem a intenção de interferir (ou colaborar) com o processo (*workflow*) que irá produzir os resultados desejados.

Na literatura existem três principais métodos para recomendar atividades em *workflow*: i) baseados em frequência de ocorrência; ii) baseados em proveniência de informação; e iii) baseados em dependência de informação. O método baseado em frequência considera o número de ocorrências de uma atividade (*A*) após outra (*B*), dessa forma, são recomendadas as atividades mais frequentes sempre que *B* for inserida.

Os métodos baseados em proveniência de informação utilizam a informação histórica para recomendar atividades. A informação pode ser: dos serviços, dos usuários, da confiança entre usuários e serviços, da execução dos *workflows* ou da sua modelagem (construção).

Os métodos baseados em dependência de informação são os mais simples. Para recomendar uma atividade são considerados os tipos de dados que a atividade utiliza como entrada e suas pré-condições de execução. São recomendadas todas as atividades que satisfaçam essas condições (tipicamente chamadas de dependência de dados e de controle).

Este artigo sumariza os resultados obtidos com a revisão sistemática que busca identificar o estado da arte dos métodos de recomendação de atividades em *workflows*. As próximas seções deste artigo descrevem a metodologia utilizada na revisão sistemática, um resumo das técnicas encontradas, as conclusões e, por fim, as considerações finais.

METODOLOGIA

A revisão sistemática foi elaborada em duas etapas. A primeira foi um estudo exploratório, para compreender o tema, encontrar palavras-chave e entender termos específicos da área.

A segunda etapa foi uma revisão sistemática que seguiu a metodologia definida por Biolchini, Mian, Conte, Travassos e Horta (2007), cujo objetivo foi responder a pergunta: "Quais métodos são utilizados para recomendar atividades em *workflows*?". A revisão sistemática é dividida em três fases: planejamento, condução e extração de dados.

Planejamento

Nesta fase, foram definidos: a pergunta que a revisão pretende responder, as bibliotecas digitais utilizadas na pesquisa, a *string* de busca e os critérios de inclusão e exclusão.

A revisão sistemática pretende responder a pergunta: "Quais são as técnicas existentes para recomendar atividades em *workflows*?". Para responder a esta pergunta foram selecionadas as bibliotecas digitais: ACMDL (ACMDL, 2015), IEEExplore (IEEE, 2015) e Science Direct (DIRECT, 2015) por serem consideradas as principais bibliotecas digitais da área que disponibilizam artigos completos de diversos periódicos e conferências.

Utilizando a metodologia PICO (Huang, Lin & Demner-Fushman, 2015) foram definidos:

1. **Population**: *scientific*, *workflow* e *pipeline;*
2. **Intervention***: recommendation, provenance, suggestion, forecast, advice, design, visualization , recommender, construc, proposal , guidance, counsel, composition, activity, shimming, inference, reuse, reusing, semiautomatically, similarity, match , matching, complete, auto;*
3. **Control:** O alvo da pesquisa são as técnicas usadas;
4. **Output:** O alvo da pesquisa é descobrir como são validadas as técnicas propostas;

Dessa forma, foi obtida a *string* de busca:

(scientific **and** (workflow **or** pipeline)) **and** (recommendation **or** provenance **or** suggestion **or** forecast **or** advice **or** design **or** visualization **or** recommender **or** construct **or** proposal **or** guidance **or** counsel **or** composition **or** activity **or** shimming **or** inference **or** reuse **or** reusing **or** semiautomatically **or** similarity **or** match **or** matching **or** complete **or** auto)

Para cada resultado da pesquisa, em cada base de dados, foram lidos os resumos e conclusões de todos os artigos e aplicados os critérios de inclusão: i) artigos que utilizam técnica de recomendação de atividades em *workflows*, e exclusão: i) trabalhos não disponibilizados na integra; ii) trabalhos que não descrevem o método utilizado; e iii) trabalhos que não são da área de recomendação de atividades em *workflows*. Para classificar um artigo como "selecionado para leitura integral", ele deve satisfazer o critério de inclusão e não satisfazer nenhum dos critérios de exclusão.

Condução

Executando as *strings* de busca, as bibliotecas digitais retornaram 897 artigos. Deste total, 386 oriundos da *ACMDL*, 369 oriundos da *IEEE* e 142 oriundos da *Science Direct*. O resumo de cada artigo foi lido e os critérios de inclusão e exclusão aplicados. Após a aplicação dos critérios, foram selecionados para a leitura integral 21 artigos. A Figura 1 sumariza o resultado da aplicação dos critérios de inclusão e exclusão dos artigos.

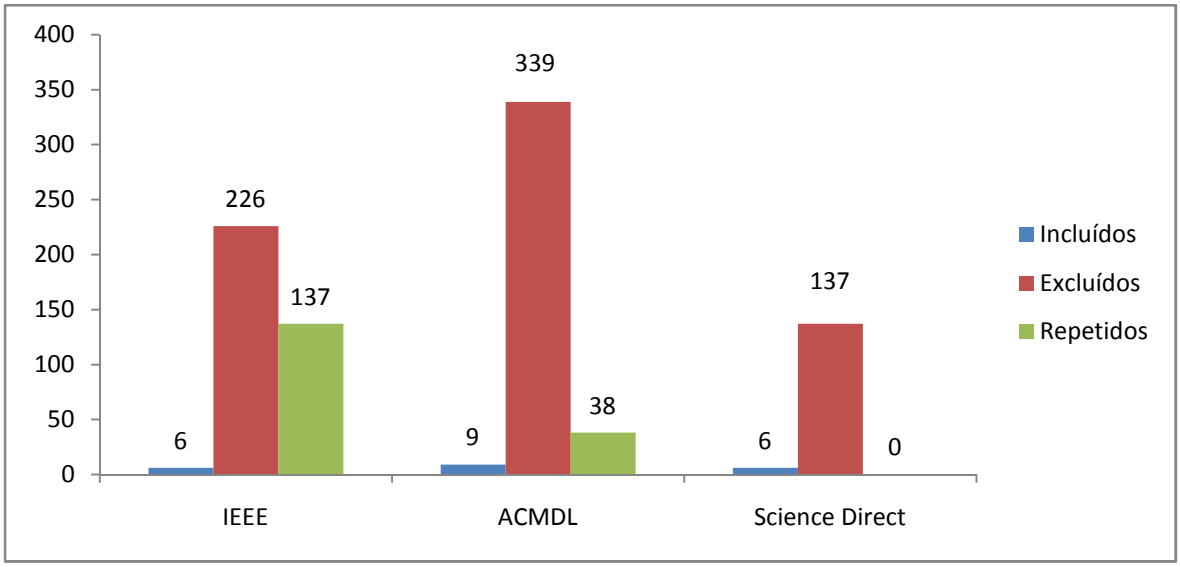

Figura 1: Resultados da revisão sistemática.

Extração de Dados

Os dados extraídos de cada artigo foram: a técnica de recomendação e metodologia de validação que podem ser vistos na Tabela 1.

Tabela 1: Dados extraídos dos artigos selecionados.

Referência	Técnica	Validação
Telea e van Wijk (1999)	Frequência	Elaborado um estudo de caso
Bomfim e Strauch (2005)	Ontologias	Elaborado um estudo de caso
Zhang (2006)	Entrada/ saída e semântica	Elaborado dois estudos de caso
Shao, Kinsy e Chen (2007)	Minera proveniência de execução	Elaborado um estudo de caso
Koop, Scheidegger, Callahan, Freire e Silva (2008)	Frequência e proveniência de execução e modelagem	Utilizados 2875 *workflows* e suas proveniências de execuções
Oliveira, Murta, Werner e Mattoso (2008)	Proveniência de execução e modelagem	Elaborado um estudo de caso
Wang, Han, Yan, Chen e Guang (2008)	Entrada e saída de atividades e semântica	Elaborado um estudo de caso
Shao, Sun e Chen (2009)	Proveniência de execução	Elaborado um estudo de caso

Wang, Cao e Li (2009)	*Itemsets* e proveniência de execução	Elaborado um estudo de caso
Yan, El-Gayyar e Cremers (2010)	Planejador	Elaborado um estudo de caso
Oliveira (2010)	Frequência	Comparado com o trabalho de Koop (2008) e os mesmos dados
Cerezo e Montagnat (2011)	Entrada/saída e semântica	Elaborado um estudo de caso
Tan, Zhang, Madduri, Foster e De Roure (2011)	Proveniência de execução	Usados *workflows* do *myExperiment* (Roure, 2014)
Zhang, Tan, Alexander, Foster e Madduri (2011)	Frequência	Usados *workflows* do *myExperiment* (Roure, 2014)
Cao et al. (2012)	Proveniência de modelagem	Elaborado um estudo de caso com dados fictícios
Diamantini, Potena e Storti (2012)	Proveniência de modelagem	Usados *workflows* do *myExperiment* (Roure, 2014)
Yao et al. (2012)	Baseado em confiança	Usados *workflows* do *myExperiment* (Roure, 2014)
Garijo et al. (2013)	Frequência e proveniência de execução	22 *workflows* e suas proveniências
Yeo e Abidi (2013)	Proveniência de execução	Elaborado um estudo de caso com dados fictícios
Bergmann e Gil (2014)	Ontologias (Contagem de arestas)	Proposta comparada com a recomendação por Levenshtein (Deza; Deza, 2009)
Zhang et al. (2014)	Frequência e semântica de metadados.	Utiliza interfaces do Programmableweb (2015)

A Figura 2 é a relação entre artigos selecionados para leitura integral e o ano de sua publicação.

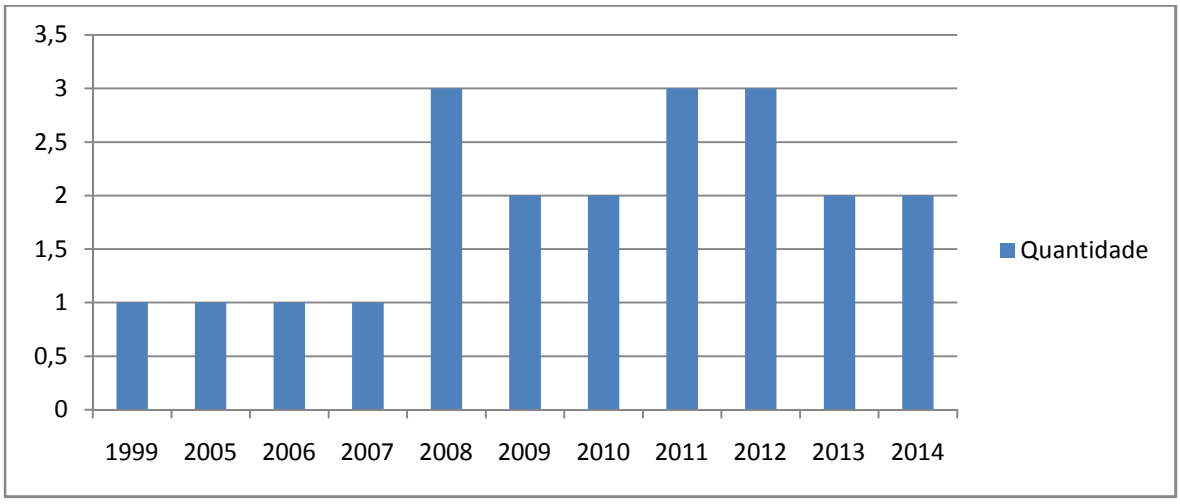

Figura 2: Número de artigos por ano de publicação.

RESULTADOS

Nesta seção são apresentadas as técnicas (para compor ou recomendar atividades) empregadas ou propostas pelos trabalhos selecionados, bem como o domínio de aplicação utilizado.

O sistema *Smartlink* proposto por Telea e Van Wijk (1999) modela os *workflows* científicos como grafos, onde as arestas correspondem ao fluxo de dados e os nós às atividades. É criado um grafo de atividades mais utilizadas e elaborada uma busca em profundidade procurando por atividades. A recomendação do *Smartlink* é baseada no grafo de atividades mais utilizadas, o que permite minimizar os seguintes problemas: i) Como conectar duas atividades?; e ii) Quais atividades podem ser conectadas a uma atividade específica?

Bomfim e Strauch (2005) desenvolveram um sistema de recomendação de atividades em *workflows* científicos baseado em ontologias. É criada uma ontologia de domínio que é utilizada para recomendar atividades em *workflows*. O autor não considera as dependências de dados e de controle para recomendar atividades. Além disso, a qualidade da recomendação não foi testada.

Zhang (2006) desenvolveu uma técnica para recomendar atividades no sistema Kepler baseada em validação ontológica de dados. Foram desenvolvidas ontologias de domínio e de tipo de dados. Durante a construção do *workflow* são validadas as atividade semântica e sintaticamente (entrada e saída). Segundo o autor, essa técnica pode melhorar a usabilidade do sistema Kepler recomendando atividades.

Shao et. al. (2007) e Shao et. al. (2009) propõem minerar a proveniência de execução e encontrar os experimentos que terminam em estado de sucesso. As execuções dos *workflows* são modeladas como grafos acíclicos dirigidos, cada execução parte do estado inicial até atingir um dos estados: i) teste; ii) não finalizado; iii) irrelevante, que não auxilia a atingir o estado de sucesso; e iv) sucesso. São considerados críticos todos os caminhos que partem do estado inicial e terminam no estado sucesso. Não foram realizados experimentos sobre recomendação, apenas sobre tempo de execução para minerar a proveniência. O autor cita que essa técnica poderia ser utilizada para recomendar os caminhos de sucesso para *workflows* em tempo de construção.

Koop et. al. (2008) recomendam *subworkflows* frequentes considerando a estrutura do *workflow*. Para tal, são encontradas (utilizando a proveniência de execução) todas as sequências de atividades posteriores ao nó âncora (aquele que vem antes do item a ser recomendado) essas serão recomendadas ao usuário

Para testar foram utilizados 2875 *workflows* e suas proveniências de execuções, geradas por estudantes durante um curso de visualização de dados, o conjunto de dados foi dividido em: i) treino, responsável por gerar caminhos; e ii) testes, responsável por usar caminhos gerados e recomendar.

Oliveira et. al. (2008) propõem recomendar trechos de *workflows* baseado em filtro colaborativo sobre a proveniência de execução e de modelagem de outros *workflows*. Sempre que uma atividade é adicionada, o sistema verifica quais as atividades seguintes foram utilizadas, através da proveniência, recomendando-as.

Na área de recomendação de serviços, Wang et. al. (2008) recomendam serviços baseados nos fatores: dependência entre serviços, modelos existentes e ordem de execução dos serviços. Considere um modelo de *workflow* no qual um serviço *b* invoca um serviço *c* e o serviço *c* invoca o serviço *d*, nessa situação um *workflow* em construção, após incluir o nó *b*, receberá as recomendações *c* e *d* ordenadas pela proximidade com *b*.

Wang, Cao e Li (2009) recomendam atividades por meio de *itemsets* frequentes minerados a partir das mudanças ocorridas nos *workflows*, cada mudança é denominada Δ, uma série destas transforma um *workflow* em outro e consiste na sequência: Δ_0, Δ_1, Δ_2 ... Δ_N. É aplicado o algoritmo *Apriori* em todas as sequências de operações Δ, dessa forma, são obtidas as regras de associação que podem ser usadas como recomendação.

Yan, El-Gayyar e Cremers (2010) propõem um planejador que procura por termos de uma ontologia, primeiramente o grafo acíclico dirigido, que representa os serviços *web*, são modelados como uma quíntupla (P, P_o, G, A, Γ) onde *P* é o conjunto de proposições, P_o é o estado inicial, *G* é o estado a ser atingido, *A* é o conjunto de ações que transformam uma proposição em outra e *Γ* é a função que transforma proposições.

No grafo os serviços serão as ações *A*, as entradas e saídas de todos os serviços serão as proposições *P*, a entrada passada pelo usuário será o estado inicial P_o e a saída esperada pelo usuário será o estado final a ser atingido *G*. O planejador começa adicionando os estados que satisfaçam as entradas das proposições existentes e que tenham uma similaridade semântica mínima, a qual é calculada por meio de graus de similaridade comparando as anotações semânticas feitas em todos os serviços com termos controlados por uma ontologia *Γ*.

A similaridade entre conceitos (c_1e c_2) da ontologia é calculada com as seguintes regras: i) c_1e c_2 são equivalentes, então é denominada exata; ii) c_2 é superconceito de c_1; iii) c_1 é super conceito de c_2; iv) são inexatos. Ao término do algoritmo é encontrado um caminho entre o estado inicial e o final que é recomendado.

Oliveira (2010) recomenda atividades de *workflows* científicos utilizando mineração sequencial. Essa abordagem permite utilizar uma modificação do algoritmo *Preorder Linked WAP* (PLWAP) desenvolvido por Ezeife, Lu e Liu (2005) para recomendar atividades. São determinadas as sequências máximas (aquelas que não estão presentes em outras sequências de um mesmo *workflow*). As sequências são a entrada para o PLWAP que define os caminhos padrões (que são usados como recomendação).

Cerezo e Montagnat (2011) elaborou um sistema que permite construir *workflows* em alto nível, utilizando ontologias de domínio, essa modelagem é convertida para uma linguagem que pode ser executada por sistemas gerenciadores de *workflow* científico. Durante a tradução, que é semi-automática, são recomendados para o usuário padrões que possuem entradas e saídas compatíveis sintaticamente e cuja similaridade ontológica é alta em relação ao *workflow* de alto nível modelado.

Tan et. al. (2011) constroem dois grafos: i) *workflows* e seus serviços, representados por nós, e as arestas representam a relação de inclusão de um serviço dentro de um *workflow*; e ii) entre operações, onde os nós representam operações dentro de serviços e as arestas operações entre serviços. Com esses grafos é possível usar o algoritmo *Apriori* para descobrir quais serviços são utilizados em conjunto por quais usuários e assim gerar recomendações.

Zhang et. al. (2011) constroem redes sociais de: i) *workflows* e serviços; ii) serviços e serviços; e iii) pessoas e serviços que permitem avaliar quais serviços são utilizados por quais *workflows*, com qual frequência e quem são os autores, o sistema foi testado com *workflows* do repositórios *myExperiment* (Roure, 2014) com validação cruzada, são recomendados os serviços mais frequentemente utilizados em *workflows* distintos.

No contexto de processos de negócio, Cao et. al. (2012) aplicam recomendação baseada em grafos durante a construção dos *workflows*. Os grafos prontos são minerados para definir padrões (sequências frequentes). Então, é calculada a distância entre os padrões e o processo de negócio (*workflow*) em construção. Os padrões com menor distância em relação ao processo de negócio em construção são recomendados ao usuário. A distância é calculada utilizando uma métrica elaborada pelos autores que considera a posição do nó no modelo pronto e no subgrafo em construção.

Diamantini, Potena e Storti (2012) recomendam fragmentos de *workflows*, modelados como um grafo dirigido acíclico, encontrando as menores subestruturas mais representativas de cada grafo. Para tal, empregam um algoritmo de agrupamento da biblioteca SUBDUE que permite reduzir os nós do grafo utilizando a métrica *Minimum Description Length* (MDL)

$$MDL = \frac{\mathrm{DL(S)} + \mathrm{DL(G|S)}}{\mathrm{DL(G)}}$$

onde *DL(S)* é a *Description Length* (DL), que é uma função para computar o número de bits necessários para representar a matriz de adjacência do subgrafo *S*, *DL(G)* é a *Description Length* do grafo original *G* e *DL(G|S) Description Length* de *G* comprimido por *S*. São recomendados os padrões mais representativos ordenados pelo valor da MDL.

Para teste foi utilizado um subconjunto de 564 *workflows* do repositório *myExperiment* (ROURE, 2014) obtendo como saída uma hierarquia de agrupamentos similares que poderiam ser usados para recomendar segundo os autores.

Yao et al. (2012) recomendam baseado em confiabilidade de serviços e autores a *ReputationNet* é um sistema de recomendação de atividades para *workflows* que considera a reputação do autor (escolaridade, especialidade e número de citações) e a popularidade dos serviços (frequência relativa). Os serviços mais populares dos autores mais confiáveis são recomendados.

Garijo et al. (2013) mineram a proveniência de execução para encontrar fragmentos frequentes de *workflows* e recomendá-los, os rastros de proveniência são representados como grafos dirigidos acíclicos. Dado um repositório de *workflows* e sua proveniência de execução o objetivo é encontrar: i) conjuntos de atividades frequentes; e ii) *subworkflows* frequentes. Utilizando o algoritmo de agrupamento do SUBDUE baseado na equação MDL. Os testes foram feitos com 22 *workflows* e suas proveniências. O resultado de estruturas frequentes encontradas foi comparado com os resultados de uma pesquisa manual (feita por um usuário). A diferença entre esta proposta e Diamantini, Potena e Storti (2012) é que esta considera a proveniência de execução para recomendar, o outro apenas os *workflows* prontos.

Yeo e Abidi (2013) adaptam a técnica de rastro de casualidade para *workflows* científicos, originalmente esta técnica é usada em *workflows* de negócios. Para isto, o autor armazena a informação de fluxo dos dados junto com o grafo de casualidade para *workflows* científicos

$$G =< N, DP, L_B, L_A >$$

onde N é o número de atividades, DP é o conjunto de caminhos do fluxo de dados, $(A, b) \in L_B$ é o conjunto de atividades para trás, tal que a execução de b é sempre realizada após alguma atividade do conjunto A e $(a, B) \in L_A$conjunto de nós para frente, tal que a é sempre executada antes de alguma das atividades do conjunto B.

Utilizando o rastro é possível criar um vetor de distâncias entre o nó âncora (que será alvo da recomendação) e os possíveis próximos nós, oriundos de L_A, esse vetor booleano contém o valor um representando a presença de uma atividade do rastro e zero caso contrário. Os vetores são gerados para todos os rastros da base de dados e suas similaridades são calculadas por meio da similaridade do cosseno (Deza & Deza, 2009).

Bergmann e Gil (2014) propõem uma ontologia para anotar as atividades, arestas, dados de cada *workflow*. A similaridade entre anotações é definida como a diferença do nível hierárquico entre elas em uma ontologia, a similaridade de atividades funciona em duas etapas: i) se as atividades são de tipos (por exemplo, algoritmo de Inteligência Artificial (IA), renderização, leitura de arquivos) diferentes sua similaridade é zero; ii) se forem do mesmo tipo, seguem a regra de similaridade de suas anotações semânticas.

A similaridade da aresta é definida como a soma entre a similaridade das atividades que ela une e a sua própria, ao comparar duas arestas de tipos diferentes, sua similaridade é zero. A similaridade do *workflow* (parcial ou total [de todos os elementos do *workflow*]) é definida como a soma de similaridades de aresta e nós sobre o número de arestas mais nós, dessa forma é possível comparar *workflows* e recomendar os mais similares.

Zhang et. al. (2014) constroem uma rede social de *workflows* e serviços (nós) e suas possíveis relações (arestas) são inclusão ou de autoria. Essa rede pode ser modelada como uma matriz Q onde $q_{i,j} = 1$ indica a inclusão de serviço em *workflow* ou como a matriz $S = Q^T Q$ onde $s_{i,j}$ representa o número de *workflows* onde os serviços (i, j) são chamados.

Quanto mais vezes dois serviços são utilizados em *workflows*, maior o grau de ligação entre os mesmos, todos os serviços são publicados com metadados, dessa forma, os autores calculam o *Term Frequency-Inverse Category Frequency* (TF-ICF) nos metadados que descrevem os serviços e suas categorias. Com base nos valores de TF-ICF cada serviço é classificado em *k* categorias. Durante a construção do *workflow* são sugeridos serviços de acordo com a métrica *Rank-Biased Overlap* (RBO), descrita no artigo. São recomendados os serviços que possuem maior RBO em relação ao *workflow* em construção.

A Figura 3 relaciona as técnicas existentes e o número de artigos estudados que as utilizam.

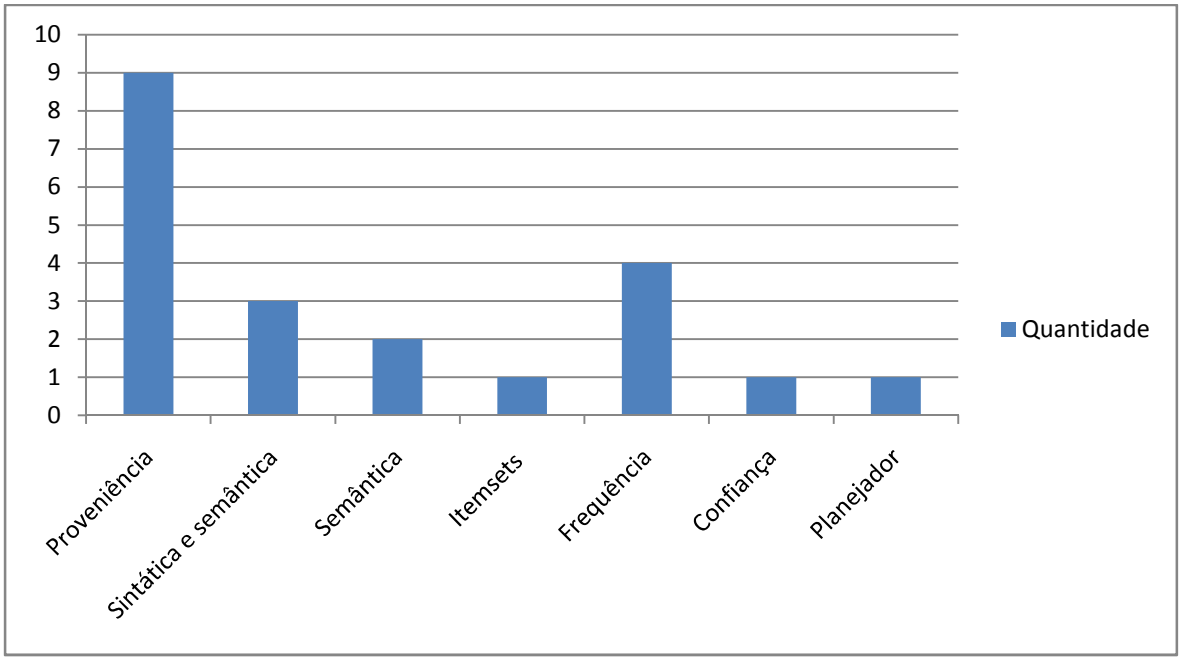

Figura 3: Número de técnicas encontradas para recomendar atividades.

CONCLUSÕES

Nesta seção são discutidos os resultados da revisão por meio da análise dos resultados presentes nas Figuras 1, 2, 3, na Tabela 1 e nos resumos elaborados na seção *Resultados*. É possível notar a existência de uma tendência no uso de técnicas baseadas em proveniência de dados, frequência e dependência da informação.

A técnica baseada em proveniência de dados (mais utilizada na literatura) tem como vantagem considerar *n* possíveis dados históricos sobre um mesmo padrão de atividade. Por exemplo, para recomendar uma atividade em um *workflow* que contenha a atividade *x*, são considerados todos os *workflows* que contenham *x* e suas atividades posteriores, a atividade com maior frequência é recomendada. Essa abordagem permite minimizar o efeito de *outliers*. Como desvantagem, possui a necessidade de uma base de dados históricos relevantes, caso contrário, *outliers* podem afetar o desempenho.

A técnica baseada em frequência tem como vantagem a simplicidade na implementação e como principal desvantagem a necessidade de uma base de dados com pouca esparsidade no uso de atividades.

A técnica baseada em dependência de informação tem como principal vantagem a facilidade de implementação. Como desvantagem, ela não leva em consideração a semântica dos dados das atividades. Por exemplo, uma *string* que representa o nome de uma espécie de bactéria é considerada similar a uma *string* que representa um CEP.

Outra tendência observada é sobre a validação dos resultados. Não há uma metodologia amplamente utilizada entre os trabalhos analisados para validação, muitos autores apenas executam a solução uma vez para “mostrar” que sua solução funciona. Não ocorrem testes com dados sintéticos ou reais. O que pode ser verificado na Tabela 1 onde doze artigos, marcados como “Elaborado um estudo de caso”, estão nessa situação.

De fato, as únicas validações encontradas que contém testes efetivos são as que utilizam o *log* do sistema. Os autores de técnicas baseadas em similaridade entre grafos afirmam utilizar um conjunto de dados sintéticos. Entretanto, nenhum dos autores descreve como o conjunto de dados foi gerado.

CONSIDERAÇÕES FINAIS

Nas últimas décadas, a computação está cada vez mais envolvida com as diversas áreas da ciência. Esse envolvimento aumenta a necessidade de partilhar recursos como: repositórios, código fonte e experimentos. *Workflows* científicos permitem esse compartilhamento, além de possibilitarem a usuários (sem conhecimento profundo em computação) desenvolver *software*. Para auxiliar neste processo de desenvolvimento (minimizando a necessidade de o usuário conhecer muitas atividades) foram propostas diversas técnicas de recomendação.

Este artigo constatou que as técnicas mais utilizadas para recomendar ou compor *workflows* são: i) baseadas em proveniência de dados; ii) baseadas em frequência; e iii) baseadas em dependência de informação. Alguns tópicos para futuras pesquisas são: técnicas híbridas usando ontologias e dependência de informação, formalização de uma metodologia para testes de qualidade em recomendação de atividades e, finalmente, a formalização da geração de conjuntos de dados para teste.

REFERÊNCIAS BIBLIOGRÁFICAS

Acmdl. *ACM Digital Library*. Recuperado em dezembro de 2014 de http://dl.acm.org.

Ayadi, N. Y., & Lacroix, Z. *Resolving Scientific Service Interoperability With Schema Mapping*. In: 2007 IEEE 7th International Symposium on BioInformaticsand BioEngineering. IEEE, 2007. p. 448--455. Recuperado em janeiro de 2014 de http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4375600.

Bergmann, R., & Gil, Y. *Similarity assessment and efficient retrieval of semantic workflows*. Information Systems, v. 40, p. 115—127. Recuperado em 12 de janeiro de 2014 de http://www.sciencedirect.com/science/article/pii/S0306437912001020.

Biolchini, J. C. A., Mian, P. G., Conte, A. C. C., Travassos, T. U. & Horta, G. *Scientific research ontology to support systematic review in software engineering.* Adv. Eng. Inform. Amsterdam, p. 133-151. Recuperado em 12 de janeiro de 2014 de http://www.sciencedirect.com/science/article/pii/S147403460600070X

Bomfim, E., Oliveira, J., de Souza, J.M. & Strauch, J. *Thoth: improving experiences reuses in the scientific environment through workflow management system.* In: Computer Supported Cooperative Work in Design, 2005. Proceedings of the Ninth International Conference on.[S.l.: s.n.], 2005. v. 2, p. 1164--1170 Vol. 2. Recuperado em 12 de janeiro de 2014 de http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=1504261

Bhagat, J., Tanoh, F., Nzuobontane E., Laurent T., Orlowski J., Roos M., Wolstencroft K., Aleksejevs S., Stevens R., Pettifer S., Lopez R. & Goble C. A. *BioCatalogue: a universal catalogue of web services for the life sciences.* Recuperado em junho 2014 de http://dx.doi.org/10.1093/nar/gkq394

Cao, Bin, Yin, Jianwei, Deng, Shuiguang, Wang, Dongjing & Wu, Zhaohui. *Graph-based workflow recommendation: on improving business processmodeling*. In: Proceedings of the 21st ACM international conference on Information andknowledge management. ACM, 2012. (CIKM'12), p. 1527--1531. Recuperado em junho 2014 de http://doi.acm.org/10.1145/2396761.2398466.

Cerezo, N. & Montagnat, J. *Scientific Workflow Reuse Through Conceptual Workflows on the Virtual Imaging Platform*. In: Proceedings of the 6th Workshop on Workflows in Support of Large-scale Science. ACM, 2011. (fWORKSg '11), p. 1--10. Recuperado em junho 2014 de http://doi.acm.org/10.1145/2110497.2110499.

Deza, M. M. & Deza, E. *Encyclopedia of Distances*. 2ed. ed. [S.l.]: Springer Berlin Heidelberg, 2009.

Diamantini, C., Potena, D. & Storti, E. *Mining Usage Patterns from a Repositoryof Scientific Workflows*. In: Proceedings of the 27th Annual {ACM} Symposium on Applied Computing. ACM, 2012. (SAC '12), p. 152--157. Recuperado em junho 2014 de http://doi.acm.org/10.1145/2245276.2245307.

Direct, Science. Science Direct. Recuperado em dezembro de 2014 de http://www.sciencedirect.com/.

Ezeife, C. I., Lu Y. & Liu Y. PLWAP *sequential mining: open source code*. In Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations (OSDM '05). ACM, New York, NY, USA, 26-35. Recuperado em junho 2014 de http://doi.acm.org/10.1145/1133905.1133910

Garijo, D., Alper, P., Belhajjame, K., Corcho, O., Gil, Y. & Goble, C. *Common motifs in scientific workflows: An empirical analysis*. In: 2012 IEEE 8th International Conference on EScience.IEEE, 2012. p. 1--8. Recuperado em junho 2014 de http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber = 6404427.

Huang, X., Lin, J. & Demner-Fushman. D. *Evaluation of PICO as a Knowledge Representation for Clinical Questions*. Recuperado em junho 2014 de http:// www.ncbi.nlm.nih.gov/pmc/articles/PMC1839740/.

Ieee. *IEEE Xplore Digital Library*. Recuperado em junho 2014 de http://ieeexplore.ieee.org/Xplore/home.jsp.

Koop, D., Scheidegger, C.E., Callahan, S.P., Freire, J. & Silva, C.T. *VisComplete: Automating Suggestions for Visualization Pipelines* In Visualization and Computer Graphics, IEEE Transactions on , vol.14, no.6, pp.1691,1698, Recuperado em junho 2014 de http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4658192.

Oliveira, Frederico Tosta de. *UM SISTEMA DE RECOMENDAÇÃO PARA COMPOSIÇÃO DE WORKFLOWS*. 2010. 91 f. Dissertação (Mestrado) - Curso de Engenharia de Sistemas e Computação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2010.

Oliveira, F. T., Murta L., Werner C. & Mattoso M. *Using provenance to improve workflow design*. In: Provenance and Annotation of Data and Processes.Springer Berlin Heidelberg, 2008, (Lecture Notes in Computer Science, v. 5272).p. 136--143. Recuperado em junho 2014 de http://link.springer.com/chapter/10.1007%2F978-3-540-89965-5_15.

Programmableweb. *Programmableweb***.** Recuperado em dezembro de 2014 de http://www.programmableweb.com/.

Roure, C. G. D. D. myExperiment. Recuperado em dezembro de 2014 de http://www.myexperiment.org/.

Shao, Q.,Kinsy, M., & Chen, Y. *Storing and Discovering Critical Workflows from Log in Scientific Exploration*. In: 2007 IEEE Congresson Services (Services 2007). Recuperado em junho 2014 de http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4278799.

Shao, Q., Sun, P., & Chen, Y. *Efficiently discovering critical workflows in scientific explorations. Future Generation Computer Systems*, v. 25, n. 5, p. 577—585. Recuperado em junho 2014 de http://www.sciencedirect.com/science/article/pii/S0167739X08000897.

Tan W., Zhang J., Madduri, R., Foster, I. & De Roure, D. *Providing Map and GPS Assistance to Service Composition in Bioinformatics*. In: 2011 IEEE International Conference on ServicesComputing. IEEE, 2011.p. 632--639. Recuperado em junho 2014 de http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6009313.

Telea, A. & VanWijk, J. J. *Vission: An Object Oriented Dataflow System for Simulation and Visualization.* Recuperado em junho 2014 de http://link.springer.com/chapter/10.1007/978-3-7091-6803-5_21.

Yan L., El-Gayyar & Cremers, A. B. *Semantics Enhanced Composition Planner for Distributed Resources*, Distributed Computing and Applications to Business Engineering and Science (DCABES), 2010 Ninth International Symposium on , vol., no., pp.61,65, 10-12. Recuperado em junho 2014 de http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5573302.

Yao J., Tan W., Nepal, S., Shiping Chen, Zhang J., De Roure, D. & Goble, C. *Reputationnet*: *A reputation engine to enhance servicemap by recommending trusted services*. In: Services Computing (SCC), 2012 IEEE NinthInternational Conference on. [S.l.: s.n.], 2012.p. 454--461.

Yeo, P. & ABIDI, S. S. R. *Dataflow Oriented Similarity Matching for Scientific Workflows*. In: 2013 IEEE International Symposium on Parallel & DistributedProcessing, Workshops and Phd Forum. IEEE, 2013. p. 2091--2100. Recuperado em junho 2014 de http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6651115.

Wang, J., Han, Y., Yan, S., Chen W. & Guang, J. *Vinca4science*: *A personal workflow system for e-science*. In: Internet Computing in Science and Engineering, 2008. ICICSE '08. International Conference on.[S.l.: s.n.], 2008. p. 444--451. Recuperado em junho 2014 de http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4548305.

Wang, Y., Cao, J. & Li, M. *Change Sequence Mining in Context-Aware Scientific Workflow*. In: 2009 IEEE International Symposium on Parallel and Distributed Processingwith Applications. IEEE, 2009. p. 635--640. Recuperado em junho 2014 de http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5207868.

Zhang, J. *Ontology-Driven Composition and Validation of Scientific Grid Workflows in Kepler: a Case Study of Hyperspectral* Image Processing, Grid and Cooperative Computing Workshops, 2006. GCCW06. Fifth International Conference on , vol., no., pp.282,289. Recuperado em junho 2014 de http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=4031563.

Zhang J., Tan W., Alexander, J., Foster, I. & Madduri, R. *Recommend-As-You-Go: A Novel Approach Supporting Services-Oriented Scientific Workflow Reuse*. In: 2011 IEEE International Conference onServices Computing. IEEE, 2011.p. 48--55. Recuperado em junho 2014 de http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6009243.

Zhang J., Lee, C., Xiao, S., Votava, P., Lee, T.J., Nemani, R. & Foster, I., *A. Community-Driven Workflow Recommendations and Reuse Infrastructure*. In: 2014 IEEE 8th International Symposium on ServiceOriented System Engineering. IEEE, 2014.p. 162--172. Recuperado em junho 2014 de http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6830902.