An ontology and frequency-based method to recommend activities in scientific workflows

Adilson Khouri*
Universidade de São Paulo
1000 Av. Arlindo Béttio
São Paulo, Brazil
adilson.khouri.usp@gmail.com

Luciano Digiampietri†
Universidade de São Paulo
1000 Av. Arlindo Béttio
São Paulo, Brazil
luciano.digiampietri@gmail.com

ABSTRACT

This paper presents a novel and hybrid approach to recommend activities in scientific workflows for datasets of activities with no provenance, no data reliability between authors and without prior semantic annotations. It is based on activities frequency and domain ontology (in this work, bioinformatics ontology). The other three main contributions from this paper are: a comparison of different techniques for activities recommendation using real workflows obtained from the myExperiment repository; the modeling of the activities recommendation problem as a classification problem, using the following classifiers: CART, Naive Bayes, Neural Network (MLP), Suport Vector Machines(SVM) and K-Nearest-Neighbor(KNN); the modeling of the activities recommendation problem as a regression problem, using the following regressors: binomial, CART, MARS, Neural Network and Suport Vector Regression (SVR).

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Recommender Systems; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Algorithms, Experimentation

Keywords

Activities Recommendation with Ontology, Scientific Workflows, Artificial Intelligence

1. INTRODUÇÃO

Uma das ferramentas para auxiliar no gerenciamento de experimentos científicos são os sistemas gerenciadores de *workflows*. *Workflows científicos* são processos estruturados e ordenados, construídos de forma manual, semi-automática ou automática que permitem solucionar problemas científicos utilizando atividades, que podem ser: i) blocos de código fonte; ii) serviços; e iii) *workflows* finalizados [29]. Estes sistemas facilitam a criação de novos experimentos, compartilhamento dos resultados e reutilização de atividades existentes.

Dentro dos sistemas gerenciadores de *workflow*, as atividades são tipicamente representadas como ícones gráficos com função *drag and drop*. Desta forma é possível construir experimentos computacionais arrastando ícones e preenchendo parâmetros de entrada. A maioria destes sistemas fornecem conjuntos de atividades básicas que podem ser utilizadas em diferentes domínios, por exemplo, uma atividade que calcula o valor médio de um conjunto de dados, é aplicável em biologia, física, astronomia e outras áreas. Porém, há uma pré-condição para se reutilizar e/ou criar *workflows*: conhecer quais são as atividades disponíveis.

Atualmente há um grande número de atividades disponíveis em repositórios como *myExperiment* que armazena mais de 2.500 *workflows* [6] e *BioCatalogue* que disponibiliza mais de 2.464 serviços [3]. O grande número de atividades e o baixo reuso de algumas atividades e *workflows* [29] motivam a construção de técnicas para recomendar atividades aos cientistas durante a composição dos *workflows*.

Sistemas de recomendação permitem aos cientistas aproveitar o poder de reutilização de workflows científicos sem a necesidade de conhecer todas as atividades ou criar atividades com mesma funcionalidade. Esses sistemas funcionam como filtro de atividades recomendando para o usuário atividades que lhe sejam úteis.

Este artigo [28] apresenta uma estratégia híbrida para recomendar atividades em workflows científicos baseada em frequência de atividades em conjunto com uma ontologia de domínio (*knowledge-base* híbrido, com MoC *dataflow*) para conjuntos de dados sem proveniência, sem dados de confiabilidade entre autores e sem anotações semânticas prévias. Além disso sugere uma modelagem do problema de recomendar atividades em workflows científicos para que seja solucionado por classificadores como: Suport Vector Machine (SVM), Naive Bayes (NB), K-Nearest-Neighbor (KNN), Classification and Regression Trees (CART) e Rede Neural (MLP). Também são utilizados os seguintes regressores como: Suport Vector Regression (SVR), CART, Rede Neural, Multivariate Adaptive Regression Splines (MARS) e regressão bi-

*Adilson Lopes Khouri.

†Luciano Antonio Digiampietri.

nomial (RB). E uma comparação das soluções da literatura correlata com as propostas.

O restante do artigo tem a seguinte estrutura na subseção 1.1 são definidos os sistemas de recomendação, seus problemas e desafios, as possíveis soluções destes. Na subseção 1.2 são apresentados os sistemas gerenciadores de workflows científicos e os workflows científicos. A subseção 1.3 apresenta os desafios de recomendar atividades em workflows científicos. A seção 2 apresenta os trabalhos da literatura correlata, a seção 3 apresenta a metodologia utilizada no trabalho, a seção 4 explica brevemente as técnicas usadas pelos classificadores e regressores a seção 5 apresenta o resultado dos experimentos realizados. Por fim a seção 6 concluí o artigo e apresenta possíveis trabalhos futuros.

1.1 Sistemas de Recomendação

Sistemas de recomendação têm como objetivo recomendar itens que sejam interessantes aos usuários, formalizando: seja C o conjunto de todos os usuários, S o conjunto de todos os itens que podem ser recomendados, u a função de utilidade que metrifica o quanto um item s é útil para um determinado usuário c, $u : C \times S \rightarrow R$ onde R é um conjunto totalmente ordenado. Para cada usuário $c \in C$ queremos escolher $s^{'} \in S$ que maximize a função de utilidade

$$\forall c \in C, \quad s_c^{'} = \arg\max_{s \in S} u(c, s) \tag{1}$$

Em sistemas de recomendação a função utilidade u não está definida para todo o espaço $C \times S$, isso força os sistemas de recomendação a extrapolar o espaço conhecido [1].

Para solucionar esse problema foram propostas diferentes técnicas para recomendar itens, as quais [9] classificam em seis grupos. A primeira, *Content-based*, recomenda itens similares a outros selecionados anteriormente pelo próprio usuário, suas limitações são: i) análise limitada do conteúdo do item que será recomendado, geralmente há falta de descrição semântica do item; ii) superespecialização quando o usuário recebe recomendações similares demais as suas escolhas; e iii) novos usuários precisam avaliar um número mínimo de itens antes que o sistema possa recomendar itens para ele.

A segunda, *Collaborative Filter*, recomenda itens que já foram selecionados por outros usuários *similares*, tem como limitações: i) o problema de novos usuários (como identificar com quem eles são similares?); ii) novos itens somente serão recomendados ao passo que forem sendo avaliados por usuários; iii) dados esparsos, alguns poucos usuários costumam avaliar muitos itens e a maioria avalia poucos itens tornando a matriz de utilidade (usuários × itens) esparsa, pois o número de avaliações feitas tende a ser muito menor do que o número de sugestões a serem realizadas. Dessa forma, itens raros (que foram avaliados por poucos) dificilmente serão recomendados.

A terceira, *Hybrid approaches*, combina características das técnicas existentes tentando minimizar suas limitações. A quarta, *Community-based*, é baseada em informações da rede social (comunidade) do usuário. A recomendação é realizada de acordo com a preferência dos colegas e amigos do usuário ao invés de preferências de desconhecidos, é um tipo de especialização do filtro colaborativo herdando suas características.

A quinta, *Demographic*, utiliza atributos como região, idade, idioma para recomendar, surgiu para tentar minimizar o problema de esparsidade e é uma especialização do filtro colaborativo considerando que usuários com mesmos dados demográficos podem ser considerados similares.

A sexta, *Knowledge-based*, recomenda itens de acordo com o domínio de aplicação, a função de similaridade estima quanto a descrição do problema é similar a solução recomendada. Tem como limitação a necessidade de descrições semânticas (usando, por exemplo, ontologias) sobre o domínio, usuário e o problema.

1.2 Sistemas Gerenciadores de Workflow Científicos

Sistemas gerenciadores de *workflows* científicos são infraestruturas de *software* que permitem a construção, reutilização, captura de proveniência e pesquisa de experimentos científicos representados na forma de *workflows* [21]. Os *workflows* possibilitam modelar e executar soluções computacionais para problemas científicos, combinando dados e operações sobre dados em estruturas configuráveis formadas por atividades [12].

Para modelar *workflows* científicos há vários paradigmas ou modelos (*Model of Computation* - [MoC]), que representam a forma como os dados são trocados entre atividades e os tipos de operações sobre dados. Neste trabalho serão abordados dois MoCs: *dataflow* e *control flow*, o primeiro, mais utilizado em *workflows* científicos, realiza transformações sobre os dados, analisa/visualiza dados e elabora simulações, o segundo, mais utilizado em *workflows* de negócio, enfatiza eventos, fluxogramas e sequências de atividades a serem desenvolvidas [20]. Atualmente é comum encontrar sistemas gerenciadores de *workflows* que combinem estes dois paradigmas.

Workflows científicos tipicamente utilizam muitas atividades do tipo *dataflow* e poucas do tipo *control flow* [20], suas atividades podem ser classificadas pela suas estruturas, são denominados como *subworkflows* quando formadas por várias atividades encadeadas e encapsuladas [22], denominadas *atividade* quando constituídas por uma única atividade [11] e *Shim* quando funcionam como adaptadores/conectores, entre duas atividades incompatíveis sintaticamente [19].

Durante as fases de construção e execução dos *workflows* alguns sistemas gerenciadores permitem a captura da proveniência dos dados, isto é, as fontes da informação utilizada, entidades, processos envolvidos na construção ou entrega de um artefato [34]. As quais, podem ser classificadas, segundo [18], nos seguintes tipos: i) *prospective provenance* que modela a especificação de um *workflow*, funcionando como uma abstração/receita do mesmo e pode ser capturada durante a construção; e ii) *retrospective provenance* que modela as execuções dos *workflows*, quais tarefas foram executadas e quais transformações sobre os dados ocorreram, esse tipo de proveniência pode ser capturado durante a execução do *workflow*.

A construção de *workflows* consiste na inclusão de diferentes tipos de atividades, na conexão destas atividades, na ligação entre os dados de entrada e as atividades bem como no preenchimento de alguns parâmetros das atividades.

Para auxiliar nessa construção foram propostas técnicas para compor automaticamente ou recomendar atividades. A primeira consiste em definir o problema a ser modelado e o sistema gerenciador conecta automaticamente as atividades construindo *workflows* para o usuário, essa técnica é recomendada para usuários que não conhecem detalhes específicos do processo e/ou não desejam se envolver nas especificidades de como o *workflow* irá resolver o problema modelado.

A segunda ocorre durante a construção manual das atividades e o sistema gerenciador sugere ao usuário algumas atividades que podem ser úteis para o *workflow* em construção. Esta sugestão geralmente é baseada em medidas de similaridade ou buscando-se atividades em *workflows* parecidos com o que está sendo desenvolvido ou buscando-se atividades usadas por usuários com o mesmo perfil do usuário atual. A técnica de recomendação é indicada para usuários mais experientes que desejam ter participação ativa na construção do *workflow*.

1.3 Recomendação em workflows científicos

Construir um sistema de recomendação para *workflows* científicos envolve duas principais tarefas, a primeira é maximizar a função de utilidade descrita pela equação (1), em outras palavras recomendar itens que satisfaçam o usuário, e a segunda, resolver problemas específicos do domínio, no caso deste projeto de mestrado, recomendar atividades para *workflows* científicos que apresentam as restrições de dependência entre entrada e saída de atividades, dependência semântica entre atividades e ordem das atividades.

A dependência entre entrada e saída de atividades implica que o tipo de dado (inteiro, *string*, *boolean*) das saídas da atividade anterior devem ser compatíveis com os tipos de dados das entradas da atividade a ser recomendada, por exemplo, a atividade A tem como saída dois inteiros e uma *string*, dessa forma qualquer outra atividade B, a ser recomendada para completar o *workflow* que contém A deve ter como entrada dados compatíveis com estes três tipos (ou com um subconjunto deles) ou será necessária a utilização de uma atividade do tipo *Shim*.

Conectar duas atividades por meio de compatibilidade de entrada e saída ou indiretamente, com uso de *Shims*, não garante que o *workflow* execute ou que o problema do usuário seja solucionado, isso ocorre em função da possível incompatibilidade semântica entre atividades, tomando novamente a atividade A, suponha que a *string* represente o nome de um gene, e a atividade a ser recomendada recebe como parâmetro uma *string* de conexão com a base de dados. Assim, estas atividades não serão compatíveis.

Além das dependências é necessário conectar as atividades na ordem correta, ao contrário de sistemas de recomendação de filmes, onde estes podem ser recomendados em ordens distintas sem afetar o resultado final da recomendação, nessa área a ordem das atividades é relevante. Por exemplo, dadas duas atividades: uma que consulte um banco de dados e outra que atualize a informação que a primeira consulta a ordem de execução destas atividades trará diferentes resultados.

Essas características motivaram a criação de técnicas específicas para recomendar atividades em *workflows* científicos, como as citadas no seção de correlatos (2), as quais consideram validação sintática (entrada e saída de atividades), frequência de uso de atividades, comparação de subgrafos, proveniência de dados, uso de semântica e *itemsets* frequentes.

2. TRABALHOS CORRELATOS

A literatura correlata apresenta algumas técnicas para recomendar atividades em workflows científicos, como usar a frequência de ocorrência de pares de atividades como proposto nos trabalhos [27, 30, 4, 10, 38, 37].

A proveniência (log de eventos) de execução/modelagem como proposto em [24, 25, 8, 16, 13, 33]. Compatibilidade entre entrada e saída de atividades do workflow como em [35, 2, 36]. Utilizando anotações suas métricas de similaridade [7, 39].

Utilizar a confiabilidade entre autores e serviços como em [32]. Itemsets frequentes com em [26, 31] utilizam o algoritmo *Apriori* para descobrir quais serviços são utilizados em conjunto por quais usuários e assim gerar recomendações.

3. METODOLOGIA

O conjunto de dados para teste foi obtido do site myExperiment [23] e contém 72 workflows de bioinformática utilizados para teste dos experimentos da literatura correlata e os propostos por este artigo. Este conjunto é uma matriz M onde as linhas são os workflows e as colunas todas as atividades disponíveis o elemento $M_{i,j} = 0$ informa que o workflow i não contém a atividade j e elemento $M_{i,j} = 1$ informa que o workflow i contém a atividade j.

Para testar as técnicas de recomendação, o conjunto de dados foi dividido em 90% para treinamento e 10% para testes. A escolha dos diferentes parâmetros foi efetuada de forma exaustiva selecionando diversos valores treinando os algoritmos e testando-os.

Para as técnicas baseadas em classificação ou regressão o conjunto de dados tem uma diferenciação, uma atividade foi removida de cada workflow e foram sugeridas outras 59 atividades (as mais frequentes do conjunto de dados) como possíveis recomendações e a atividade correta (a removida é sugerida como recomendação TRUE).

Dessa forma, são adicionadas 59 linhas na matriz M, cada uma representando uma possível recomendação, e uma nova coluna que informa se a sugestão é correta ou não. Para evitar o desbalanceamento entre exemplos positivos e negativos optou-se por adicionar mais 59 linhas com a atividade correta. Neste conjunto de dados cada conjunto de 118 linhas adicionadas representa uma lista de atividades recomendadas.

A resposta dos classificadores é binária representando o resultado final da recomendação. Enquanto que a resposta dos

regressores é uma predição numérica dos atributos de acordo com alguns modelos de regressão, nesse caso cada valor predito é utilizado como limiar, valores maiores são considerados *TRUE* e valores menores que este limiar são considerados *FALSE*. O melhor valor de limiar é selecionado por meio de testes e os valores preditos são convertidos para *TRUE* ou *FALSE*.

Cada técnica de recomendação utilizada sugere ao usuário uma lista de algumas atividades da base de dados selecionadas de acordo com algum critério seguidas por todas as outras atividades da base ordenadas alfabeticamente. Dessa forma, a atividade desejada (o alvo da recomendação) será sempre encontrada e a métrica número de acertos não será utilizada. As métricas usadas como critério de qualidade para definir o melhor sistema de recomendação serão $S@k$ e Mean Reciprocal Rank (MRR).

4. TÉCNICAS UTILIZADAS

O classificador KNN começa com um conjunto de treinamento rotulado e outro de testes não rotulado (ambos com as mesmas dimensões). Para cada instância do conjunto de teste o KNN encontra os k vizinhos mais próximos do conjunto de treinamento, a instância é classificada na classe com maior número de vizinhos próximos [17].

O classificador e/ou regressor CART tem um funcionamento padrão ambos utilizam uma estrutura em árvore para aprender e tomar decisões para tal utiliza um nó raiz, diversos nós de decisão e os nós folha. Cada dado a ser classificado inicia o fluxo de tomadas de decisão pela raiz da árvore, em cada nó de decisão é elaborado um teste lógico baseado em algum atributo a construção das árvores depende do algoritmo que está sendo utilizado e do critério de divisão aplicado. O regressor CART usa a métrica de impureza a soma residual ao quadrado $\sum_{i=1}^{n}(y_i - f(x_i))^2$ ao invés do índice de Gini (ou ganho de informação) cada nó folha será preenchido pela média dos valores dos exemplos de treinamento atribuídos naquela folha. Outros aspectos como construção, poda e treinamento são idênticos a árvores de classificação [5].

Classificadores baseados em métodos Bayesianos utilizam os dados treinados para calcular a probabilidade observada de cada classe baseada em valores de suas características. Esses classificadores são melhor aplicados em problemas onde a informação de vários atributos pode ser considerada simultaneamente para gerar uma estimativa de probabilidades de saídas. O algoritmo Naive Bayes, é um exemplo desses classificadores, é uma aplicação do teorema de Bayes adaptado para classificação, assume algumas pré-condições ingênuas sobre os dados como: i) independência de características; e ii) que todas as características são igualmente importantes. No mundo real essas pré condições são falhas mas ainda assim o algoritmo possuí um desempenho satisfatório [17].

Redes neurais tem como inspiração o cérebro humano, seus neurônios e suas conexões seu objetivo é modelar uma relação de entradas e saídas ponderadas que são definidas por vários nós de processamento. Os quais são responsáveis por calcular a soma de entradas ponderadas e repassa-los para a função de ativação que determina se um sinal será enviado (ou não) para o neurônio seguinte ou para a saída da rede [15]. Quando usados como classificador a resposta da camada de saída é a classificação da instância, quando usados como regressor pode-se utilizar os pesos da rede, treinados por algum algoritmo de treinamento como o backpropagation, usando estes valores para predizer.

Suport Vector Machines (SVM), nesse trabalho é utilizado o *C-SVM*, é uma técnica que pode ser usada para classificar dados usando um hiperplano. O objetivo é escolher a posição do hiperplano tal que permita formar partições homogêneas de dados em ambos os lados da superfície de decisão. É um problema de otimização que pode ser formulado [15] como

$$\max_{\alpha} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j k(x_i, x_j) \tag{2}$$

para as seguintes restrições

$$\sum_{i=1}^{N} \alpha_i d_i = 0 \tag{3}$$

$$0 \leq \alpha_i \leq C \tag{4}$$

$$C > 0 \tag{5}$$

onde α são os multiplicadores de Lagrange, d são as saídas esperadas do problema, x é o conjunto de dados de entrada, K é uma função de kernel e C é uma constante positiva definida pelo usuário.

O SVR Ãl' uma adapataÃğÃčo do SVM para regressão , nesse trabalho é utilizado o ϵ-*SVR*, tem um objetivo oposto ao classificador, enquanto o último tenta maximizar a margem (separando ao máximo os dados) o primeiro tem por objetivo aproximar os dados ao maximo dessa margem com uma dada tolerância para erros. Sendo formulado como um problema de otimização [15] como

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_i^* - \alpha_j \alpha_j^* K(x_i, x_j) - \sum_{i=1}^{N} \alpha_i \alpha_i^* d_i \tag{6}$$

para as seguintes restrições

$$-C \leq \alpha_i - \alpha^* \leq C \tag{7}$$

$$\sum_{i=1}^{N} \alpha_i - \alpha_i^* = 0 \tag{8}$$

onde α, α^* são os multiplicadores de Lagrange, d são as saídas esperadas do problema, x é o conjunto de dados de entrada, K é uma função de kernel e C é uma constante positiva definida pelo usuário.

O algoritmo *Multivariate Adaptive Regression Splines*(MARS) é uma generalização da regressão linear por função degrau[14]. Para tal usa segmentos lineares de funções com a seguinte estrutura:

$$(x-t)_+ = \begin{cases} x - t, \text{se } x > t \\ 0, \text{cc} \end{cases} \tag{9}$$

$$(t-x)_+ = \begin{cases} t - x, \text{se } x < t \\ 0, cc \end{cases} \tag{10}$$

A ideia é formar pares de funções espelhos para cada variável independente X_j com um nó para cada valor de $x_{i,j}$ daquela

variável. O modelo de regressão MARS tem o seguinte formato:

$$f(X) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(X) \quad (11)$$

onde $h_m(X)$ é uma função ou o produto de duas ou mais funções, os coeficientes β_m são estimados pela minimização da soma do resíduo quadrado.

A regressão Binomial pode ser modelada como um *modelo generalizado linear* que é constituído tem três componentes: i) a distribuição da variável dependente(neste caso binomial); ii) o preditor linear $\alpha + \beta X = \frac{p}{1-p}$; e iii) a função *link* que relaciona a média da distribuição com o preditor linear, no nosso caso é $g(\mu) = \log(\frac{p}{1-p})$ [14].

No caso da regressão binomial a variável dependente Y segue uma distribuição normal e a função de link e o preditor são dados pela equação (14)

$$g(\mu) = \log_e \left(\frac{\pi}{1-\pi} \right) \quad (12)$$

$$g(\mu) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \quad (13)$$

$$\log_e \left(\frac{\pi}{1-\pi} \right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \quad (14)$$

onde $\pi = \mu$ é a média de Y, p são as dimensões dos dados, β os coeficientes de regressão que serão estimados por máxima verossimilhança e X os dados.

5. EXPERIMENTOS

A Tabela 1 exibe a comparação entre as diferentes técnicas propostas neste artigo e algumas técnicas da literatura correlata.

As técnicas de classificação CART, KNN, NNET, aleatório e Apriori (proposta em [26, 31]) obtiveram um desempenho muito baixo. Os três classificadores não conseguiram convergir consequentemente não realizaram uma boa classificação. Em relação as outras duas técnicas o baixo desempenho da recomendação aleatória é em virtude da ocorrência de muitas atividades (280) fazendo com que a probabilidade de escolher a atividade correta seja pequena, o apriori não considera os seguintes fatores fundamentais: a ordem de atividades, entrada e saída destas e nem semântica de atividades obtendo um baixo desempenho.

A classificação por NNET e os regressores: Binomial, CART, MARS e NNET apresentam um resultado superior em função das métricas $S@5$ e $S@10$ apresentarem atividades corretas dentre as primeiras posições, inclusive o MRR destas é superior ao das técnicas anteriores. O uso de classificador e regressor SVM ocasionou a primeira melhoria considerável com resultados 3 vezes melhores nas métricas $S@5 = 0.428$, $S@10 \geq 0.714$ e uma melhoria considerável da métrica $MRR_{Class} = 0.2958$ e $MRR_{Reg} = 0.3149$

Entre as técnicas classicas da literatura correlata, para dados: sem proveniência; sem informações de autores e confiabilidade e sem anotações semânticas prévias. Percebe-se

Table 1: Resultados das recomendações.

Técnica	$S@5$	$S@10$	$S@100$	$S@280$	**MRR**
Classificadores					
CART	0.000	0.000	0.000	1	0.0101
KNN	0.000	0.000	0.143	1	0.0102
NAIVE	0.000	0.000	0.000	1	0.0101
NNET	0.143	0.143	0.143	1	0.1524
SVM	0.428	0.714	1.000	1	0.2958
Regressores					
Binomial	0.000	0.285	0.571	1	0.277
CART	0.000	0.285	0.428	1	0.0391
MARS	0.000	0.285	0.428	1	0.0254
NNET	0.143	0.143	0.143	1	0.1524
SVR	0.428	0.857	1.000	1	0.3149
Correlatos					
Aleatorio	0.000	0.000	0.000	1	0.0097
Apriori	0.000	0.000	0.143	1	0.0102
I/O	0.000	0.428	1.000	1	0.0562
Freq. I/O	0.428	0.714	1.000	1	0.2936
Freq. I/O Onto.	0.571	0.714	1.000	1	0.3174

uma primeira melhoria ao comparar a técnica de entrada e saída de atividades (propostas em [27, 30, 4, 10, 38, 37]) com a híbrida que considera também a frequência das mesmas, recomendar a atividade mais frequente para atividades que contenham a mesma assinatura apresenta resultados melhores. Ao acrescentar a informação ontológica sobre o tipo de workflow em que as atividades estão inseridas e ordenar a lista de recomendação das atividades mais frequentes com este critério obteve-se um resultado superior.

Isso ocorre em duas situações distintas, a primeira quando a atividade a ser recomendada é a primeira do workflow. Onde a técnica por frequência não é útil recomendando todas as possíveis atividades e a segunda onde existe uma atividade anterior e suas possíveis recomendações por frequência tem um empate (duas atividades com mesma frequência). Em ambos os casos ao utilizar uma ontologia para ordenar as atividades que são do mesmo grupo ontológico obten-se resultados melhores nessas situações.

6. CONCLUSÕES

Este artigo apresentou como constribuições uma técnica híbrida para recomendar atividades em workflows científicos baseada em frequência entrada e saída e ontologias, uma possível modelagem do problema de recomendação para ser solucionado por classificadores e regressores uma comparação entre diferentes técnicas de recomendação de atividades em workflows científicos para um mesmo conjunto de dados reais obtidos no repositório myExperiment.

Os melhores algoritmos para realizar esta tarefa são SVM e Frequência com Ontologia eles obtiveram desempenhos muito próximos, a desvantagem do SVM é o tempo de treinamento muito alto para ajustar os parâmetros enquanto que a desvantagem do experimento por frequência e ontologia é a necessidade de conhecer o domínio de aplicação para aplicar e construir uma ontologia.

Como possíveis trabalhos futuros os autores pretendem uti-

lizar outras variaÃğÃţes de SVM como *v-SVR*,ϵ-*SVR* e *v-SVM*. Classificação multi-classe e um classificador misto que utilize dados dos outros classificadores como entrada.

7. ACKNOWLEDGMENTS

Os autores agradecem a agência CAPES pelo financiamento do projeto.

8. REFERENCES

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.

[2] N. Y. Ayadi and Z. Lacroix. Resolving Scientific Service Interoperability With Schema Mapping. In *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*, pages 448–455. IEEE, Oct. 2007.

[3] J. Bhagat, F. Tanoh, E. Nzuobontane, T. Laurent, J. Orlowski, M. Roos, K. Wolstencroft, S. Aleksejevs, R. Stevens, S. Pettifer, R. Lopez, and C. Goble. Biocatalogue: a universal catalogue of web services for the life sciences, June 2014.

[4] B. Cao, J. Yin, S. Deng, D. Wang, and Z. Wu. Graph-based workflow recommendation: on improving business process modeling. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 1527–1531. ACM, 2012.

[5] S. B. Connor. *Wiley Encyclopedia of Statistics in Quality and Reliability*, chapter Perfect Sampling. Wiley, 2007.

[6] C. G. David De Roure. myexperiment, junho 2014.

[7] D. de Oliveira, L. Cunha, L. Tomaz, V. Pereira, and M. Mattoso. Using Ontologies to Support Deep Water Oil Exploration Scientific Workflows. In *2009 Congress on Services - I*, pages 364–367. IEEE, July 2009.

[8] F. de Oliveira, L. Murta, C. Werner, and M. Mattoso. Using provenance to improve workflow design. In J. Freire, D. Koop, and L. Moreau, editors, *Provenance and Annotation of Data and Processes*, volume 5272 of *Lecture Notes in Computer Science*, pages 136–143. Springer Berlin Heidelberg, 2008.

[9] F. A. P. de Paiva, J. A. F. Costa, and C. R. M. Silva. A Hierarchical Architecture for Ontology-Based Recommender Systems. In *2013 BRICS Congress on Computational Intelligence and 11th Brazilian Congress on Computational Intelligence*, pages 362–367. IEEE, Sept. 2013.

[10] C. Diamantini, D. Potena, and E. Storti. Mining Usage Patterns from a Repository of Scientific Workflows. In *Proceedings of the 27th Annual {ACM} Symposium on Applied Computing*, SAC '12, pages 152–157. ACM, 2012.

[11] D. Garijo, P. Alper, K. Belhajjame, O. Corcho, Y. Gil, and C. Goble. Common motifs in scientific workflows: An empirical analysis. In *2012 IEEE 8th International Conference on E-Science*, pages 1–8. IEEE, Oct. 2012.

[12] D. Garijo, P. Alper, K. Belhajjame, O. Corcho, Y. Gil, and C. Goble. Common motifs in scientific workflows: An empirical analysis. *Future Generation Computer Systems*, 36:338–351, July 2014.

[13] D. Garijo, O. Corcho, and Y. Gil. Detecting Common Scientific Workflow Fragments Using Templates and Execution Provenance. In *Proceedings of the Seventh International Conference on Knowledge Capture*, K-CAP '13, pages 33–40, New York, NY, USA, 2013. ACM.

[14] T. J. Hastie, R. J. Tibshirani, and J. H. Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer, New York, 2009.

[15] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 3nd edition, 2007.

[16] D. Koop, C. E. Scheidegger, S. P. Callahan, J. Freire, and C. T. Silva. VisComplete: automating suggestions for visualization pipelines. *IEEE transactions on visualization and computer graphics*, 14(6):1691–8, Jan. 2008.

[17] B. Lantz. *Machine Learning with R*. Packt Publishing, Birmingham, 1nd edition, 2013.

[18] C. Lim, S. Lu, A. Chebotko, and F. Fotouhi. Prospective and Retrospective Provenance Collection in Scientific Workflow Environments. In *2010 IEEE International Conference on Services Computing*, SCC '10, pages 449–456. IEEE, July 2010.

[19] C. Lin, S. Lu, X. Fei, D. Pai, and J. Hua. A Task Abstraction and Mapping Approach to the Shimming Problem in Scientific Workflows. In *2009 IEEE International Conference on Services Computing*, SCC '09, pages 284–291. IEEE Computer Society, 2009.

[20] B. Ludascher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao, and Y. Zhao. Scientific workflow management and the kepler system. *Concurrency and Computation: Practice and Experience*, 18(10):1039–1065, 2006.

[21] T. McPhillips, S. Bowers, D. Zinn, and B. Ludäscher. Scientific workflow design for mere mortals. *Future Generation Computer Systems*, 25(5):541–551, May 2009.

[22] C. B. Medeiros, J. Perez-Alcazar, L. Digiampietri, G. Z. Pastorello Jr., A. Santanche, R. S. Torres, E. Madeira, and E. Bacarin. {WOODSS} and the Web: Annotating and Reusing Scientific Workflows. {*SIGMOD*} *Rec.*, 34(3):18–23, 2005.

[23] C. G. Roure. myexperiment, 2015.

[24] Q. Shao, M. Kinsy, and Y. Chen. Storing and Discovering Critical Workflows from Log in Scientific Exploration. In *2007 IEEE Congress on Services (Services 2007)*, pages 209–212. IEEE, July 2007.

[25] Q. Shao, P. Sun, and Y. Chen. Efficiently discovering critical workflows in scientific explorations. *Future Generation Computer Systems*, 25(5):577–585, May 2009.

[26] W. Tan, J. Zhang, R. Madduri, I. Foster, D. De Roure, and C. Goble. Providing Map and GPS Assistance to Service Composition in Bioinformatics. In *2011 IEEE International Conference on Services Computing*, pages 632–639. IEEE, July 2011.

[27] A. Telea and J. J. van Wijk. vission: An object oriented dataflow system for simulation and

visualization. In *PROCEEDINGS OF IEEE VISSYM*, pages 95–104, 1999.

[28] M. Uschold and M. King. Towards a methodology for building ontologies. In *In Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95*, 1995.

[29] F. Wang, H. Deng, L. Guo, and K. Ji. A Survey on Scientific-Workflow Techniques for E-science in Astronomy. In *2010 International Forum on Information Technology and Applications*, volume 1, pages 417–420. IEEE, July 2010.

[30] J. Wang, Y. Han, S. Yan, W. Chen, and G. Ji. Vinca4science: A personal workflow system for e-science. In *Internet Computing in Science and Engineering, 2008. ICICSE '08. International Conference on*, pages 444–451, 2008.

[31] Y. Wang, J. Cao, and M. Li. Change Sequence Mining in Context-Aware Scientific Workflow. In *2009 IEEE International Symposium on Parallel and Distributed Processing with Applications*, pages 635–640. IEEE, 2009.

[32] J. Yao, W. Tan, S. Nepal, S. Chen, J. Zhang, D. De Roure, and C. Goble. Reputationnet: A reputation engine to enhance servicemap by recommending trusted services. In *Services Computing (SCC), 2012 IEEE Ninth International Conference on*, pages 454–461, 2012.

[33] P. Yeo and S. S. R. Abidi. Dataflow Oriented Similarity Matching for Scientific Workflows. In *2013 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum*, pages 2091–2100. IEEE, May 2013.

[34] R. Zeng, X. He, and W. van der Aalst. A Method to Mine Workflows from Provenance for Assisting Scientific Workflow Composition. In *2011 IEEE World Congress on Services*, pages 169–175. IEEE, July 2011.

[35] J. Zhang. Ontology-driven composition and validation of scientific grid workflows in kepler: a case study of hyperspectral image processing. In *Proceedings of the Fifth International Conference on Grid and Cooperative Computing Workshops*, GCCW '06, pages 282–289. IEEE Computer Society, 2006.

[36] J. Zhang. Ontology-Driven Composition and Validation of Scientific Grid Workflows in Kepler: a Case Study of Hyperspectral Image Processing. In *2006 Fifth International Conference on Grid and Cooperative Computing Workshops*, pages 282–289. IEEE, 2006.

[37] J. Zhang, C. Lee, S. Xiao, P. Votava, T. J. Lee, R. Nemani, and I. Foster. A Community-Driven Workflow Recommendations and Reuse Infrastructure. In *2014 IEEE 8th International Symposium on Service Oriented System Engineering*, pages 162–172. IEEE, Apr. 2014.

[38] J. Zhang, W. Tan, J. Alexander, I. Foster, and R. Madduri. Recommend-As-You-Go: A Novel Approach Supporting Services-Oriented Scientific Workflow Reuse. In *2011 IEEE International Conference on Services Computing*, pages 48–55. IEEE, July 2011.

[39] L. Zhang, Y. Y. Y. Y. Wang, P. Xuan, A. Duvall, J. Lowe, A. Subramanian, P. K. Srimani, F. Luo, and Y. Duan. Sesame: A new bioinformatics semantic workflow design system. In *2013 IEEE International Conference on Bioinformatics and Biomedicine*, pages 504–508. IEEE, Dec. 2013.