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

Adilson L. Khouri[1,*] Luciano A. Digiampietri[2]
1 School of Arts, Sciences and Humanities, University of São Paulo, Brazil
2 School of Arts, Sciences and Humanities, University of São Paulo, Brazil
∗ **E-mail: Corresponding adilson.khouri.usp@gmail.com**

Abstract

The number of activities provided by scientific workflow management systems is large, which requires scientists to know many of them to take advantage of the reusability of these systems. To minimize this problem, the literature presents some techniques to recommend activities during the scientific workflow construction. This project specified and developed a hybrid activity recommendation system considering information on frequency, input and outputs of activities and ontological annotations. Additionally, this project presents a modeling of activities recommendation as a classification problem, tested using 5 classifiers; 5 regressors; a SVM classifier, which uses the results of other classifiers and regressors to recommend; and Rotation Forest , an ensemble of classifiers. The proposed technique was compared to other related techniques and to classifiers and regressors, using 10-fold-cross-validation, achieving a MRR at least 70% greater than those obtained by other techniques.

Introduction

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 [**?**]. 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* [**?**] e *BioCatalogue* que disponibiliza mais de 2.464 serviços [**?**]. O grande número de atividades e o baixo reuso de algumas atividades e *workflows* [**?**] 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 [**?**] 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 binomial (RB). E uma comparação das soluções da literatura correlata com as propostas.

O restante do artigo tem a seguinte estrutura na subseção **??** são definidos os sistemas de recomendação, seus problemas e desafios, as possíveis soluções destes. Na subseção **??** são apresentados os sistemas gerenciadores de workflows científicos e os workflows científicos. A subseção **??** apresenta os desafios de recomendar atividades em workflows científicos. A seção **??** apresenta os trabalhos da literatura correlata, a seção **??** apresenta a metodologia utilizada no trabalho, a seção

?? explica brevemente as técnicas usadas pelos classificadores e regressores a seção **??** apresenta o resultado dos experimentos realizados. Por fim a seção **??** concluí o artigo e apresenta possíveis trabalhos futuros.

Materials and Methods

Os *workflows* foram obtidos no repositrio *myExperiment* [**?**], por meio do *software wget* [**?**]. Aps efetuar o *download* dos 2481 *workflows* em formato *xml*, foi utilizado o analisador de cdigo *Beautiful Soup* [**?**], para organizar o conjunto de dados em uma base de dados relacional.

Os dados foram usados em trs estruturas distintas, um grafo usado para as tcnicas da literatura correlata que consideram a ordem, uma matriz simples usada para as tcnicas que no usam ordem das atividades. Uma matriz adaptada para modelar o problema como um problema de classificao (binria) e regresso.

Grafo

Matriz simples

Os *workflows* da rea de bioinformtica (totalizando 73) em conjunto com suas atividades (totalizando 280) foram convertidos em uma matriz $M_{i,j}$ em que cada linha i representa um *workflow*, cada coluna j representa uma das 280 atividades e cada clula da matriz M representa a existncia $M_{i,j} = 1$, ou no $M_{i,j} = 0$, da atividade da coluna j no *workflow* i. A tabela 1 apresenta um exemplo, fictcio, de matriz M. Para a realizao dos testes, para cada linha da tabela 1 removida uma atividade e recomendada uma lista de possveis atividades. O objetivo do sistema de recomendao identificar corretamente qual a atividade est faltando no workflow (isto , aquela que foi removida).

Table 1. Exemplo de matriz de entrada.

Workflow	**Ativ 01**	**Ativ 02**	...	**Ativ 280**
01	1	0	...	0
02	1	1	...	1
03	1	0	...	1
⋮	⋮	⋮	⋮	⋮
73	1	0	...	0

Matriz adaptada

Para usar tcnicas de classificao e regresso foram propostas algumas alteraes no conjunto de dados original, descrito na tabela 1, as quais podem ser visualizadas na tabela 2. Cada *workflow* foi replicado 118 vezes.

Destes, 59 so uma cpia idntica ao original, enquanto que dos outros 59 foi removida uma mesma atividade para todos os *workflows*, e foi adicionada uma nova atividade representando uma possvel recomendao. Dessa forma, para cada *workflow* original haver 59 instncias corretas e 59 instncias incorretas e este tipo de informao ser utilizada para treinar os classificadores ou regressores.

Table 2. Exemplo de matriz de entrada para tcnicas de classificao e regresso

#	*Workflow*	Ativ 01	Ativ 02	...	Ativ 279	Ativ 280	Rtulo
1	01	1	0	...	0	0	T
2	01	1	0	...	0	0	T
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
59	01	1	0	...	0	0	T
1	01	0 (removida)	1 (adicionada)	...	1	0	F
2	01	0 (removida)	0	...	1 (adicionada)	0	F
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
59	01	0 (removida)	0	...	0	1 (adicionada)	F
	⋮						
1	73	1	1	...	0	0	T
2	73	1	1	...	0	0	T
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
59	73	1	1	...	0	0	T
1	73	1 (adicionada)	0 (removida)	...	1	0	F
2	73	1	0 (removida)	...	1 (adicionada)	0	F
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
59	73	1	0 (removida)	...	0	1 (adicionada)	F

A escolha de 59 atividades a serem recomendadas foi feita por duas razes. A primeira selecionar as 59 atividades com maior frequncia na base de dados. A segunda a limitao computacional: replicar as 280 possveis recomendaes poderia ser invivel em termos de treinamento. Foram replicadas 59 instncias de *workflows* idnticas consideradas corretas, isto com a atividade correta no removida, para garantir o balanceamento entre classes. A ltima alterao foi adicionar uma coluna indicando se a recomendao da atividade proposta a correta, isto , a pertencente ao respectivo *workflow* (T) ou no (F).

Data Storage & Enrichment

s

Analysis

s

Results

A tabela 3 exibe os resultados de cada sistema recomendador usado. As tcnicas que possuem a letra C em subscrito so classificadores; as que possuem letra R em subscrito so regressores; e as que no tem nada so da literatura correlata. Cada sistema efetua suas recomendaes de acordo com seus diferentes critrios em uma lista inicial. Em seguida, as atividades no recomendadas so acrescentadas ao final da lista inicial. Dessa forma, a atividade correta sempre ser encontrada, e o fator que diferencia os sistemas de recomendao a posio em que as atividades ocupam na lista de atividades final que contm 280 posies.

O sistema baseado em *Aleatoriedade* no precisou de treinamento. O algoritmo apenas selecionava aleatoriamente as atividades formando uma lista de atividades recomendadas. Esse sistema recomendou menos de 3% das atividades corretas entre as dez primeiras posies. A maioria das atividades corretas foram classificadas prximas a posio 140 que a posio mdia das listas recomendadas. Os valores das mtricas

Table 3. Resultados dos sistemas de recomendao

#	**Tcnica**	**S@1**	**S@5**	**S@10**	**S@50**	**S@100**	**S@280**	**MRR**
1	Aleatrio	0,0037	0,0260	0,0280	0,0300	0,0400	1,0000	0.033
2	*Apriori*	0,0037	0,0385	0,0559	0,0568	0,0570	1,0000	0,037
3	KNN_C	0,0037	0,0685	0,0959	0,5068	1,0000	1,0000	0,040
4	Rede neural_C	0,0137	0,1507	0,1781	0,8082	1,0000	1,0000	0,089
5	CART_C	0,0274	0,1233	0,3699	0,7671	1,0000	1,0000	0,113
6	CART_R	0,1370	0,1370	0,2603	0,6164	1,0000	1,0000	0,114
7	Naive Bayes_C	0,0274	0,1507	0,3425	0,6301	1,0000	1,0000	0,114
8	Binomial_R	0,0822	0,1918	0,2055	0,8493	1,0000	1,0000	0,136
9	Rede neural_R	0,1096	0,2603	0,2603	0,2603	1,0000	1,0000	0,154
10	MARS_R	0,1233	0,2055	0,2192	0,7260	1,0000	1,0000	0,167
11	SVM_R	0,1233	0,3151	0,4932	0,8493	1,0000	1,0000	0,238
12	FES	0,1474	0,2603	0,3699	0,8671	1,0000	1,0000	0,196
13	SVM_C	0,2425	0,4658	0,4932	0,7123	1,0000	1,0000	0,244
14	SVM composto_C	0,2515	0,4458	0,5232	0,7623	1,0000	1,0000	0,314
15	Rotation Forest_C	0,2925	0,4558	0,5432	0,7723	1,0000	1,0000	0,324
16	FESO	0,3425	0,4658	0,5932	0,8123	1,0000	1,0000	0,334

$S@280 = 1$ e $S@100 = 0,0400$ indicam que a maior parte dos itens corretos foi encontrado aps a centsima posio. Esse sistema foi proposto como um marco de comparao.

O sistema que usa a tcnica *Apriori* obteve seu melhor desempenho quando os parmetros *confiana* e *suporte* foram definidos como *sem limitao*, isto , no foi estabelecido um valor de confiana ou suporte mnimo para considerar possveis regras de associao criadas. Todas as regras foram consideradas vlidas. Mesmo sem restringir esses valores, os resultados desse sistema foram superiores apenas ao sistema baseado em Aleatoriedade. Recomendando menos de 6% das atividades corretas entre as 50 primeiras posies, sua preciso ainda baixa com valor de $MRR = 0,037$. Os baixos resultados dessa tcnica acontecem devido ao fato de desconsiderar a ordem das atividades durante a gerao das regras e, consequentemente, da recomendao.

O sistema baseado em *KNN* foi treinado para diferentes valores do parmetro $k = [1 : 100]$ que representa o nmero de vizinhos mais prximos (de acordo com a distncia Euclidiana) que sero considerados para classificar. Este sistema apresentou os melhores resultados de recomendao para o valor de $k = 2$. Mesmo assim, menos de 10% dos itens corretos foram encontrados entre as dez primeiras posies da lista e 50% dos itens entre os 50 primeiros itens. De acordo com a mtrica MRR, a posio mdia dos itens recomendados foi distante da primeira posio da lista $MRR = 0,040$. Esses resultados indicam que classificar atividades de acordo com a distncia entre grupos de vizinhos prximos no uma abordagem adequada para o problema.

O sistema que usa uma rede neural MLP como classificador teve uma melhoria de quase quatro vezes na mtrica $S@1$ de $0,0037$ para $0,0137$ em relao ao *KNN*. Para o treinamento da rede foram usados os parmetros: i) nmero de neurnios η (variando entre $1 : 40$); ii) taxa de aprendizagem α (variando entre $10^{-7} : 10^{+1}$); iii) duas camadas escondidas; e iv) arquitetura totalmente conectada. Os melhores resultados de classificao foram obtidos para $\eta = 18$ e $\alpha = 10^{-4}$ obtendo 17% de itens classificados entre as dez primeiras posies da lista, e 80% entre as 50 primeiras posies, o que representa uma melhoria de 30% em relao a tcnica *KNN*. O valor da mtrica $MRR = 0,089$ apresentou uma taxa duas vezes mais elevada que a do *KNN*, esse aumento de preciso indica que o poder de generalizar da rede neural para solucionar problemas no lineares foi mais eficiente que a capacidade de generalizao das tcnicas anteriores.

O sistema baseado em CART como classificador, que tem como caracterstica tratar dados categricos,

apresentou um resultado superior ao da rede neural. O treinamento usou os parmetros: i) valor mnimo de diviso $\gamma = [0 : 30]$; ii) tamanho mximo da rvore final $\delta = [0 : 10000]$; iii) valor mnimo de variao para realizar uma diviso $cp = [10^{-7} : 10^{+1}]$; iv) funo de diviso (ξ) como ndice de Gini ou ganho de informao. O melhor resultado foi para $gamma = 0$, $\delta = 30$, $cp = 10^{-3}$ e $\xi =$ Ganho de informao.

Os resultados desse sistema foram aproximadamente duas vezes melhores que os da rede neural. Isso indica uma tendncia de bons resultados para tcnicas que lidem com dados categricos por natureza. Essa melhoria indicou um aumento de 26% na mtrica MRR que representa um aumento da preciso do sistema, alm disso posicionou 13% dos itens procurados na primeira posio e 26% nas primeiras 50 posies.

O sistema baseado em CART como regressor, teve seu melhor valor com os parmetros $gamma = 2$, $\delta = 20$, $cp = 10^{-5}$ e $\xi =$ Ganho de informao. A recomendao que usou valores contnuos apresentou um resultado superior ao $CART_C$ nas mtricas $S@1$ e $S@5$ e um resultado inferior para $S@10$ e $S@50$, e a preciso geral (MRR) do $CART_R$ foi levemente superior.

O sistema baseado no classificador Naive Bayes obteve resultados muito prximos ao do regressor CART. O treinamento ocorreu modificando o atributo *correo de Laplace* com valores entre $[0 : 100]$. O melhor resultado ocorreu para o valor zero obtendo 34% dos itens recomendados entre as dez primeiras posies e 63% entre as 50 primeiras posies. Em contrapartida, o valor de MRR no sofreu grande variao.

O sistema baseado em regressor binomial apresentou melhoria em relao ao Naive Bayes e rede neural (tcnicas que apresentaram resultados prximos). O treinamento dessa tcnica ocorre por mxima verossimilhana de um modelo generalizado linear aproximado por uma distribuio binomial. Os resultados para $S@5$ e $S@50$ foram superiores que das tcnicas anteriores e o valor da mtrica MRR melhorou em aproximadamente 19% em relao a tcnica Naive Bayes. Isto indica que aproximar a varivel dependente por uma distribuio binomial e estimar seus parmetros por verossimilhana uma ideia potencialmente interessante para tratar este problema.

A rede neural como regressor, que utiliza o peso da rede neural como sada, foi treinada de forma anloga rede neural usada como classificador. O melhor resultado foi obtido para os valores de $\eta = 10$ e $\alpha = 10^{-2}$ recomendando 26% dos itens corretos entre as dez primeiras posies da lista. A preciso do sistema (MRR) melhorou 13% em relao ao regressor binomial. Esses resultados indicam que usar um regressor ao invs de um classificador apresenta um resultado melhor para esse tipo de problema, quando solucionado com redes neurais.

O sistema que usou o algoritmo MARS como regressor apresentou um resultado superior rede neural (usada como regressor) em $12,5\%$ na mtrica $S@1$, trs vezes mais atividades recomendadas entre as 50 primeiras e um aumento de preciso geral (MRR) de 8%. Esse resultado mostra que as curvas criadas pelas diversas funes conectadas do MARS obtiveram uma generalizao melhor que da rede neural. O treinamento dos parmetros foi por verossimilhana.

O regressor SVM apresentou resultados duas vezes melhores que o algoritmo MARS para a medida S@10, pois em 49% das recomendaes o item correto estava entre as dez primeiras posies da lista de recomendaes. O valor de MRR tambm foi superior (42%). O treinamento foi feito por otimizao de margem com os valores de $c = [10^{-7} : 10^{2}]$, $\epsilon = [10^{-7} : 10^{2}]$, valores de tolerncia $\beta = [10^{-7} : 10^{2}]$, funes de *kernel*: i) linear; ii) sigmoide; iii) polinomial; e iv) radial, os parmetros do *kernel* polinomial so: i) $p = [1 : 10]$ que a potncia da funo. Os melhores valores encontrados foram para $c = 1$, $\epsilon = 1$, $\beta = 10^{-4}$, *kernel* polinomial com $p = 2$. Esse resultado um indcio que o problema no linearmente separvel, pois foi usada uma funo de *kernel* polinomial para mapear o problema em alta dimenso e projet-lo novamente para uma dimenso mais baixa. Os autores acreditam que esta caracterstica foi responsvel pelo bom desempenho desse regressor.

Dentre os sistemas propostos pela literatura, o sistema baseado em entrada, sada e frequncia (FES) [**?**] o que apresenta os melhores resultados. Nos experimentos realizados, este sistema identificou o item correto entre as dez primeiras posies da lista de recomendao em 37% dos casos, e obteve um valor de $MRR = 0,196$.

O sistema baseado no algoritmo SVM para classificao foi o nico classificador que superou os resultados

dos regressores. Seu treinamento foi anlogo ao SVM para regresso. Sua melhor execuo foi para os valores $c = 10^{-1}$, $p = 10^{-4}$ e *kernel* linear. Esta execuo, para a mtrica $S@1$ foi 64% melhor que a da tcnica FES e o valor da preciso geral (MRR) aumentou 24%. Este resultado indica que a soluo utilizando *kernel* para mapeamento em alta dimenso uma proposta eficiente no caso de classificadores.

O sistema SVM composto, que executa sobre os resultados dos outros sistemas de recomendao, apresentou um desempenho superior ao SVM para classificao. Seu treinamento foi anlogo ao do SVM_C e seu melhor desempenho foi para os parmetros $c = 10^{-2}$, $p = 1$ e *kernel polinomial*. Houve uma melhoria de 3% na mtrica $S@1$ e 28% na mtrica MRR, essa melhoria em virtude do uso do resultado de outros classificadores em conjunto com a reduo de esparsidade do conjunto de dados.

O sistema utilizando *Rotation Forest* apresentou o segundo melhor resultado, seu treinamento utilizou os parmetros: i) valor mnimo de diviso $\gamma = [0 : 30]$; ii) tamanho mximo da rvore final $\delta = [0 : 10000]$; iii) valor mnimo de variao para realizar uma diviso $cp = [10^{-7} : 10^{+1}]$; iv) funo de diviso (ξ) como ndice de Gini e ganho de informao; v) $K = [1 : 10]$ como nmero de parties; vi) $L = [1 : 10]$ como o nmero de classificadores; e vii) valores de corte $0,25; 0,5; 0,75$. Essa melhoria foi em virtude de usar em conjunto uma tcnica de classificao do tipo *ensemble* e trs limiares de corte, os quais foram estabelecidos para converter os valores numricos (da mdia dos L classificadores) em valores binrios.

A tcnica FESO, apresentou um resultado superior s demais. Este considera o uso de frequncia, entrada e sada e informaes semnticas sobre as atividades. Em comparao com as demais tcnicas seu resultado foi superior para todas as mtricas calculadas, exceto $S@50$ para algumas tcnicas. Em relao tcnica FES, seu resultado foi superior. Em particular, parte dessa melhora justificada pelos casos em que a atividade correta teria frequncia zero no conjunto de treinamento, pois ela permite recomendar baseada na ontologia (usando as atividades que contenham a ontologia do novo *workflow*). Alm disso, para o caso em que h empate entre duas atividades com o critrio de entrada e sada e a frequncia a tcnica proposta apresenta um fator a mais para ser utilizado como desempate.

Algumas tendncias observadas com esses resultados foram que aumentar a informao sobre dados na recomendao melhora o seu desempenho, como o resultado dos experimentos: 2, 12 e 14 mostram. Uma segunda tendncia que o classificador SVM foi o nico que obteve um melhor resultado que os regressores, indicando que solues por maximizao de espao entre dados em alta dimenso podem ser uma rea de estudo promissora. Uma terceira tendncia o uso de classificadores compostos e *ensembles*, os quais apresentaram resultados promissores. No caso do *ensemble* h um indcio que tcnicas desse tipo, que usem limiares para converter os valores da mdia dos resultados do conjunto L em valores binrios, tm resultados promissores na recomendao de atividades.

Related Work

Para estabelecer o estado da arte os autores realizaram uma reviso sistemtica [**?**] cujos resultados so sumarizados na figura **??**. Que permite afirmar que h diversos estudos sobre recomendao de atividades em workflows cientficos. A maior parte destes estudos desconsidera uso de ontologias e/ou anotaes semnticas e as tcnicas mais usadas so baseadas em provenincia de informao.

Os trabalhos de **?** e **?**, que consideram a minerao sequencial de atividades como *itemsets* desconsideram a ordem das atividades e a semntica das mesmas. A proposta de **?** desconsidera apenas a semntica das atividades. Esta proposta de mestrado considera a ordem de atividades que um fator importante na recomendao conforme visto no captulo de conceitos fundamentais.

Os trabalhos de **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?**, **?** consideram a ordem das atividades, entrada e sada e provenincia dos dados. Suas limitaes so a necessidade de dados de provenincia, pois nem todo SGWC armazena essas informaes, alm de desconsiderar informao semntica dos *workflows* e atividades. Este projeto no necessita de informaes de provenincia e considera a semntica da informao por meio de uma ontologia hierarquizada e validada por um especialista da rea.

O trabalho de **?** usa apenas um mapeamento entre atividades e ontologia desconsiderando a entrada

e sada, o que potencialmente gera recomendaes ineficientes. Neste projeto so consideradas s entradas e sadas de cada atividade individualmente, alm do uso de uma ontologia de domnio.

?,**?** desconsideram o uso de semntica das atividades e da frequncia de suas ocorrncias em pares. Nesse projeto de mestrado so considerados esses dois fatores.

O trabalho de **?** exige dados que permitam calcular a confiana dos usurios e dos seus *workflows*. Repositrios como *myExperiment* [**?**] no exigem dos usurios o preenchimento de todos os seus dados, de forma que grande parte das informaes relacionadas a este aspecto no so preenchidas pelos usurios. Alm disso, os autores desconsideram a semntica das atividades e *workflows*. Este projeto de mestrado considera a semntica de *workflows* e no necessita da informao sobre a confiana dos usurios.

Os trabalhos de **?**, **?** e **?** desconsideram o uso de semntica de dados para recomendar, o que um limitante conforme discutido por **?**, **?**. No presente mestrado, a frequncia considerada em conjunto com a ontologia de domnio.

Os trabalhos de **?**, **?**, **?** desconsideram o uso de uma ontologia hierarquizada e validada por um especialista. Dessa forma, a qualidade das anotaes semnticas questionvel. Nesse projeto foi construda uma ontologia usando uma metodologia e esta foi validada por um especialista.

Os trabalhos de **?**,**?** consideram o uso de frequncia e ontologia, como neste projeto, porm recomendam *subworkflows* o que limita as recomendaes de atividades. Apenas atividades usadas em fragmentos comuns de *workflows* podero ser recomendadas. Em outras palavras, se a atividade se encontra no "meio" de um *subworkflow* esta nunca poder ser recomendada individualmente. No presente mestrado, todas as atividades tem possibilidade de ser recomendadas, mesmo que no final da lista de recomendao. Alm disso, apresenta uma recomendao mais abrangente, pois trata o caso de atividades simples, *subworkflows* e *Shims* (ver seo **??**).

Neste mestrado o problema de recomendao de atividades foi tambm modelado como um problema de classificao e regresso, usando para isso 5 classificadores; 5 regressores; um classificador SVM composto (que usa o resultado dos outros classificadores e regressores para recomendar) e um *ensemble* de classificadores (*Rotation Forest*).

A partir da figura **??** possvel notar a existncia de uma tendncia no uso de tcnicas baseadas em provenincia de dados, frequncia e dependncia da informao. A partir de 2014 a literatura comeou a considerar estratgias hbridas que usam provenincia e algum tipo de informao semntica. No ano de 2015 foram publicados dois artigos propondo estratgias hbridas para recomendar que usam frequncia e algum tipo de informao semntica para recomendar *subworkflows*.

A tcnica baseada em provenincia de dados (mais utilizada na literatura) tem como vantagem considerar diversos dados histricos sobre um mesmo padro de atividade. Por exemplo, para recomendar uma atividade em um *workflow* que contenha a atividade x, so considerados todos os *workflows* que contenham x e suas atividades posteriores, a atividade com maior frequncia recomendada. Essa abordagem permite minimizar o efeito de *outliers*. Como desvantagem, possui a necessidade de uma base de dados histricos relevantes, caso contrrio, *outliers* podem afetar o desempenho.

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

A tcnica baseada em dependncia de informao tem como principal vantagem a facilidade de implementao. Como desvantagem, ela no leva em considerao a semntica dos dados das atividades. Por exemplo, uma *string* que representa o nome de uma espcie de bactria considerada similar a uma *string* que representa um CEP.

Outra tendncia observada sobre a validao dos resultados. No h uma metodologia amplamente utilizada entre os trabalhos analisados para validao. Muitos autores apenas executam a soluo uma vez para "mostrar" que sua soluo funciona. No ocorrem testes com dados sintticos ou reais, o que pode ser verificado na tabela **??** em que 11 artigos esto nessa situao (marcados na tabela como "Elaborado um estudo de caso").

Conclusion

Este trabalho desenvolveu uma tcnica hbrida para recomendar atividades em *workflows* cientficos, que usa compatibilidade sinttica, frequncia e ontologias de domnio para recomendar atividades, denominada FESO. Alm disso, tambm modelou o problema de recomendao como um problema de regresso e classificao em inteligncia artificial.

A principal ideia do projeto foi acrescentar informaes semnticas estruturadas para o sistema de recomendao. Conforme foi apresentado no captulo de resultados (captulo **??**), esta estratgia atingiu melhores resultados do que as outras tcnicas implementadas, sendo que a medida MRR aumentou 70% em relao as outras estratgias.

Para encontrar as tcnicas da literatura correlata, foi realizada uma reviso sistemtica (captulo **??**). Nessa reviso foram encontradas as tcnicas, suas restries, suas vantagens e as formas que foram validadas. O prximo passo foi implement-las e compar-las com as solues propostas neste mestrado, incluindo as solues baseadas em classificadores e regressores.

Para realizar a comparao foi organizado um banco de dados relacional de *workflows* e suas atividades. Tambm foi necessrio estabelecer uma metodologia para comparar diferentes tcnicas de recomendao de atividades para um mesmo conjunto de dados com as mesmas mtricas de validao $S@k$ e MRR (descritas na seo **??**).

Ao comparar todas as tcnicas, foram constatados determinados aspectos do conjunto de dados, como o fato das atividades no serem independentes; o problema no ser linearmente separvel; e que tcnicas de agrupamento no se mostraram adequadas para solucionar este problema. Com exceo do SVM, regressores apresentaram solues mais precisas do que classificadores. Alm disso, adicionar informao nos sistemas de recomendao melhorou a preciso destes. A seguir sero listadas as principais contribuies deste mestrado e potenciais trabalhos futuros.

Acknowledgments

Os autores agradecem a agência CAPES pelo financiamento do projeto.

Figure Legends