Desenvolvimento de técnica para recomendar atividades em *workflows* científicos: uma abordagem baseada em ontologias.

Orientando: Adilson Lopes Khouri

Orientador: Prof. Dr. Luciano Antonio Digiampietri

17 de fevereiro de 2016

Sumário

Introdução

1. *e-Science*.
2. Sistemas Gerenciadores de Workflows Científicos.
   - Visualizar grandes quantidades de dados.
   - Cálculos matemáticos.
   - Análise de genomas.
3. Evitar escrever funções/métodos existentes.
4. Grande número de atividades.
5. Sistema para recomendar atividades.

Objetivo Geral

Este mestrado tem por objetivo especificar e implementar uma técnica de recomendação de atividades em workflows científicos que combine:

1. Ontologias.
2. Frequência de pares de atividades.
3. Entrada e saída de atividades.

Objetivos Específicos

1. Construir uma base de dados de *workflows* científicos.
2. Modelagem da recomendação de atividades como um problema de classificação/regressão.
3. Comparação entre diferentes técnicas da literatura e soluções propostas.

Conceitos Fundamentais

1. Sistemas Gerenciadores de Workflows Científicos.
2. Sistemas de recomendação.
3. Recomendação em *workflows* científicos.
4. Ontologias.
5. Recomendação baseada em bases de dados de *workflows* científicos.
6. Métricas de validação.
7. Recomendação a partir de banco de workflows.
8. Classificadores e regressores.

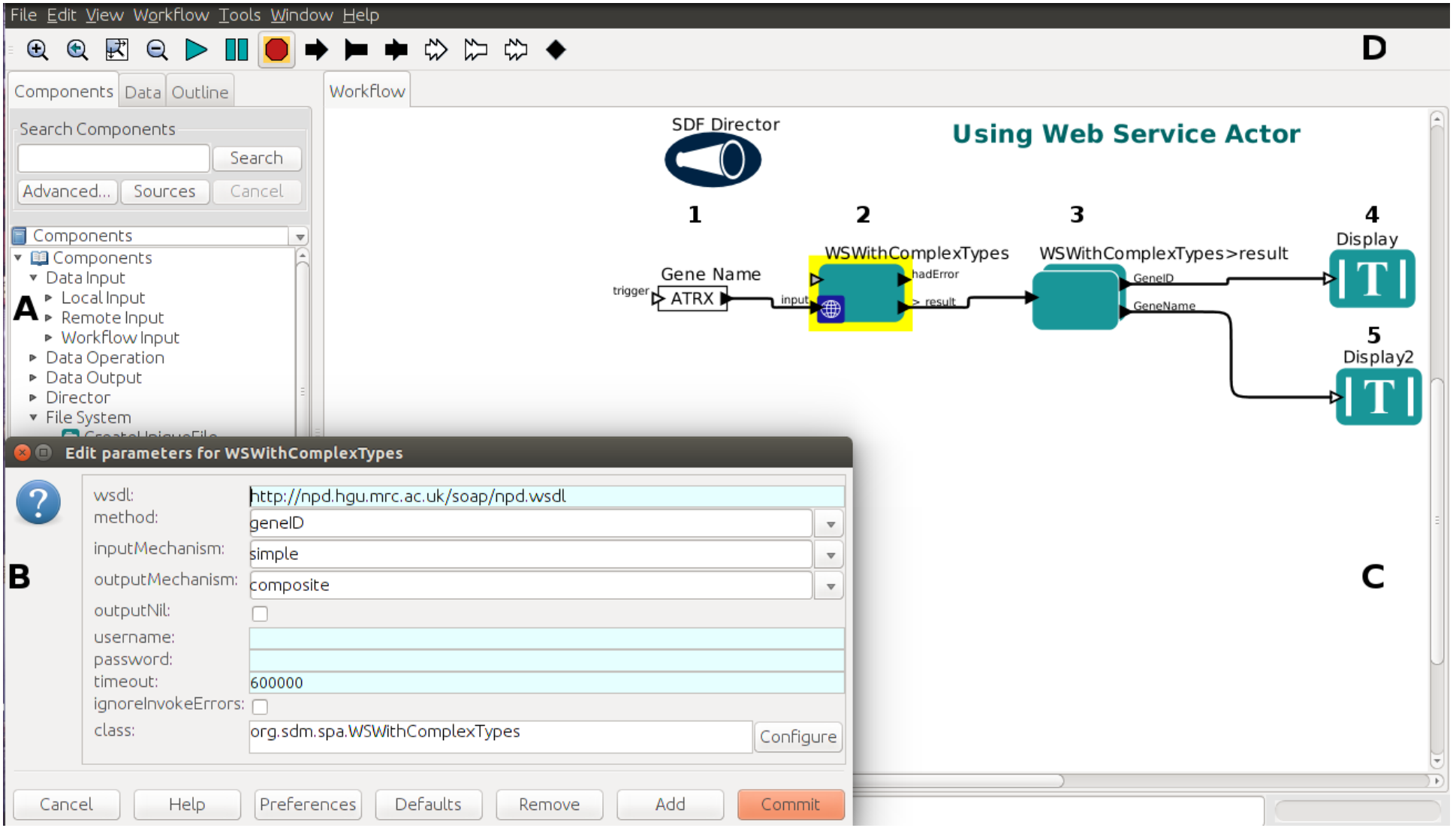

Figura: Exemplo de sistema gerenciador de *workflow* científico.

Sistemas de recomendação têm por objetivo **recomendar itens úteis** para usuários:

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

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.

Algumas estratégias utilizadas em sistemas de recomendações:

1. *Content-based*.
2. *Collaborative Filter (usuários parecidos)*.
3. *Hibrid Approach*.
4. *Community Based (usuários amigos)*.
5. *Demographic*.
6. *Knowledge-based*.

Recomendar atividades em workflows científicos exige, além da extrapolação citada, considerar as restrições:

1. Dependência entre entrada e saída de atividades.
2. Dependência semântica.
3. A ordem das atividades (citar exemplo de sistema de recomendação de música).

Ontologia é um modelo para representação de conhecimento, a qual, pode ser utilizadas para anotar semanticamente atividades.

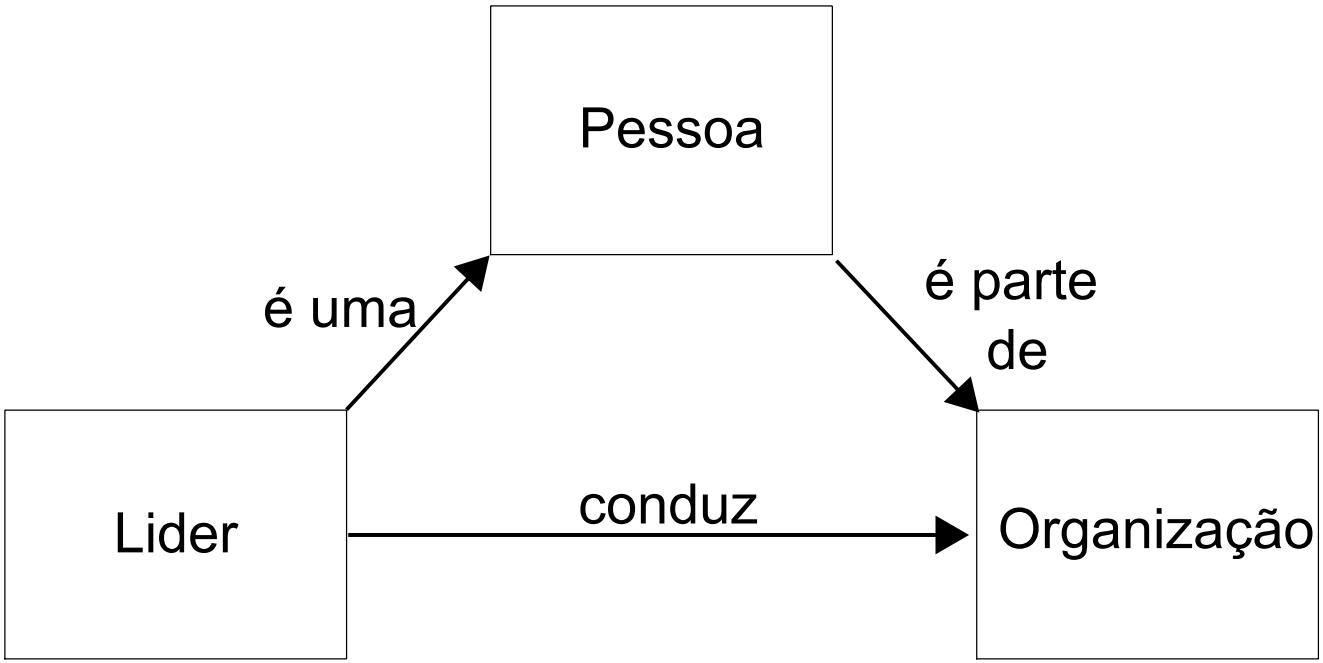

Figura: Exemplo de Ontologia

Os experimentos serão validados por *10-fold cross validation*, em cada rodada serão calculadas as métricas:

1. *Sucess at rank k* (*S@k*).
2. *Mean Reciprocal Rank* (MRR).

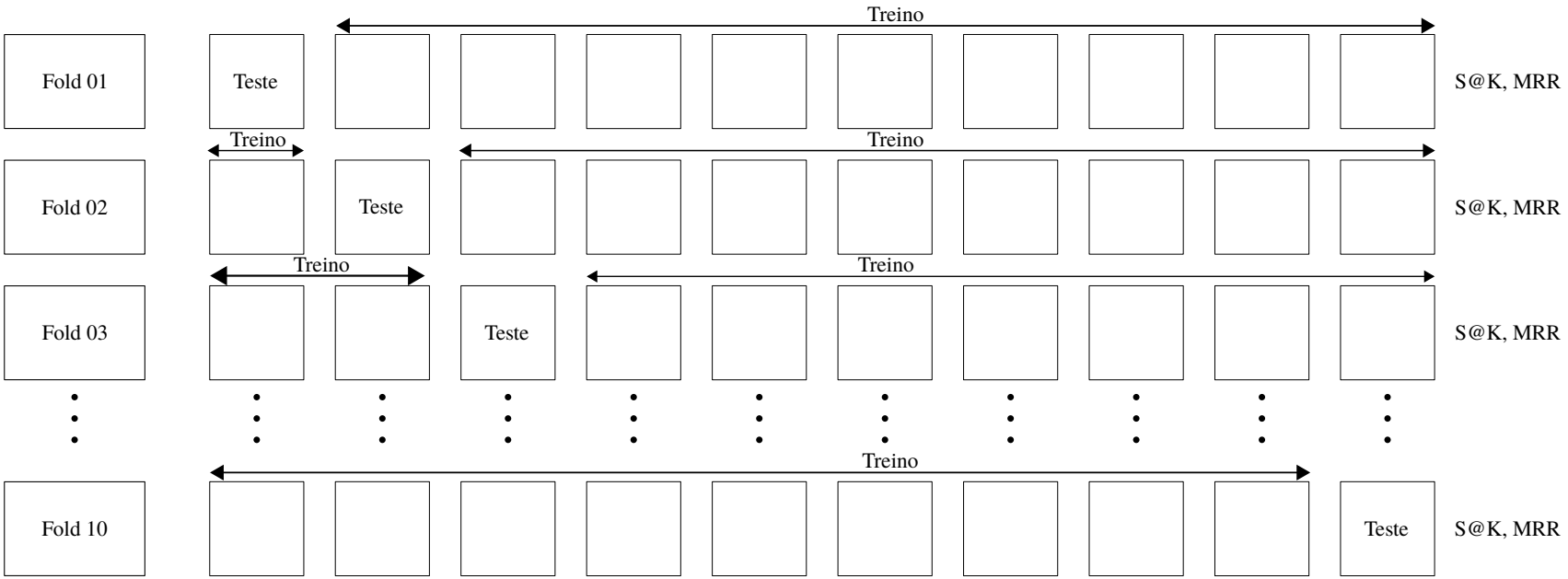

Figura: Exemplo de *10-fold cross validation*

Recomendação a partir de banco de workflows

1. Frequência;
2. *itemsets*.

Recomendação a partir de classificadores

1. CART;
2. KNN;
3. Naive Bayes;
4. Rede Neural (MLP);
5. SVM (C-SVM).

Recomendação a partir de regressores

1. CART;
2. MARS;
3. Binomial;
4. Rede Neural (MLP);
5. SVM (ϵ-SVM).

Recomendação a partir de classificadores compostos

1. SVM;
2. Rotation Forest.

Revisão Sistemática

A revisão da literatura iniciou com um estudo exploratório seguido de uma resivão sistemática. Dessa forma, foi possível:

1. Encontrar o estado da arte na área de recomendação de atividades em *workflows* científicos.
2. Compreender o problema;
3. Encontrar termos específicos da área;
4. Definir palavras-chave;

Condução

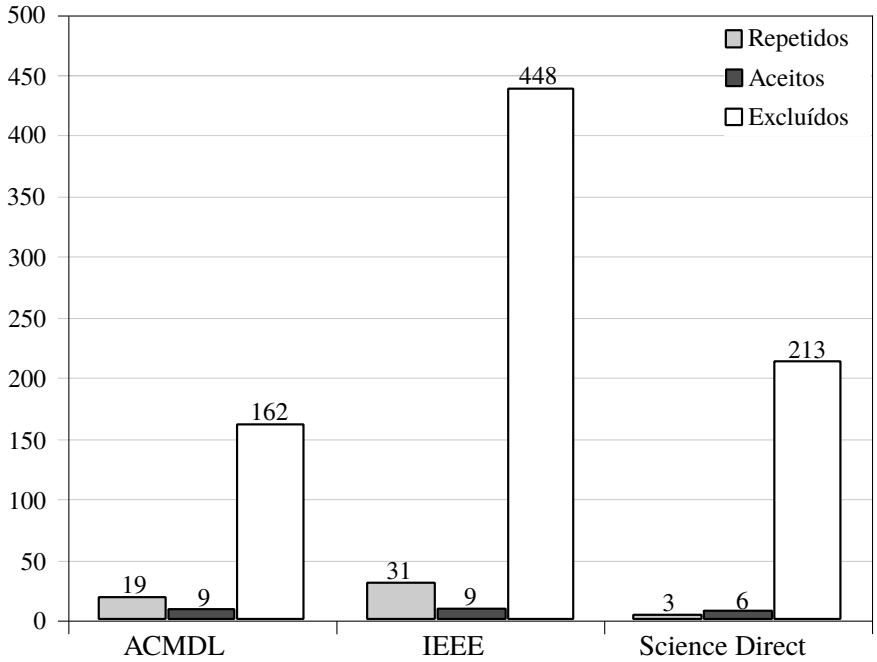

Figura: Quantidade de artigos por técnica

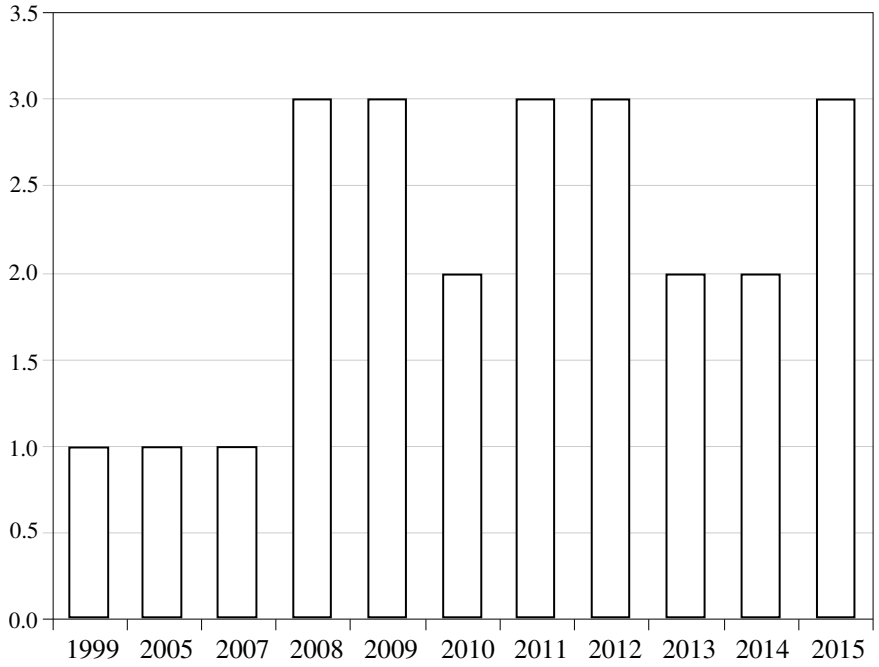

Figura: Artigos por ano de publicação

Execução

Observa-se que a técnica de proveniência é a mais usada seguida por: i) Frequência; ii) Entrada e saída; iii) Itemsets; e iv) Ontologias.

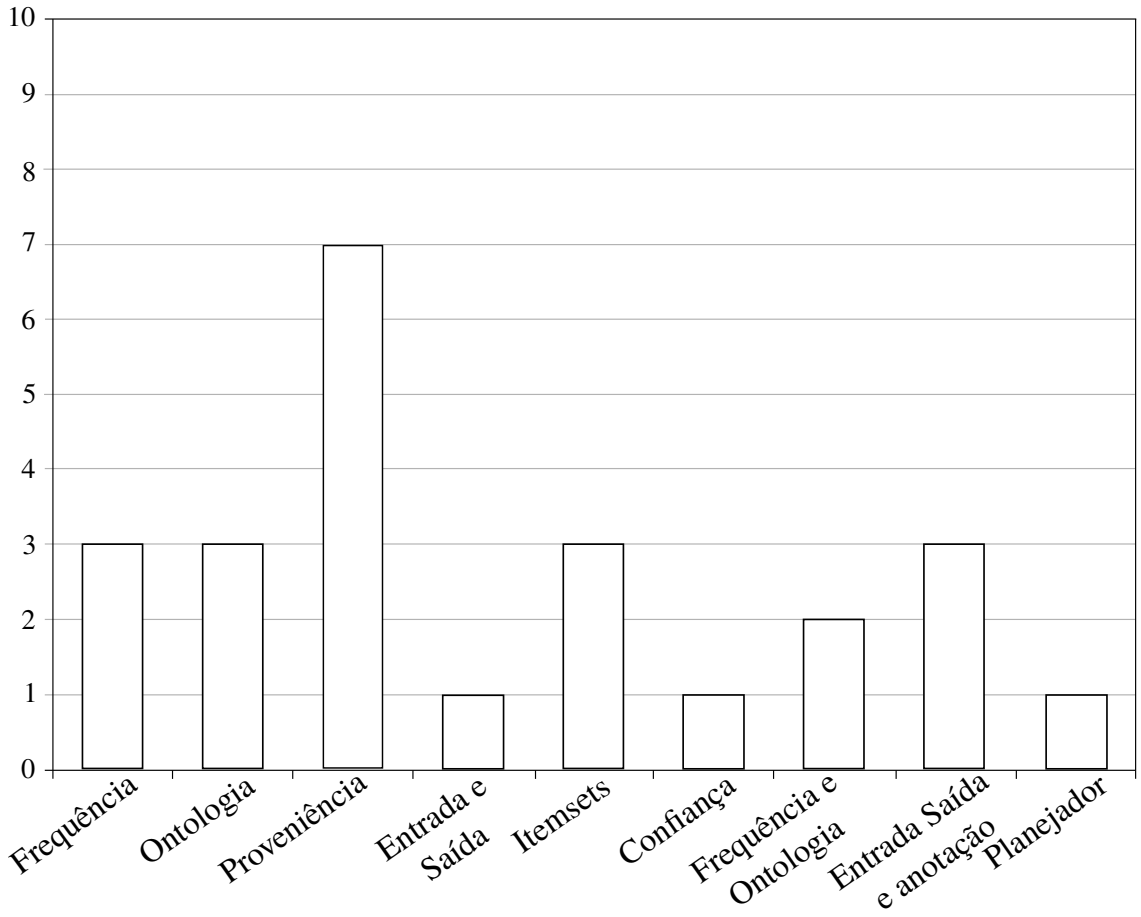

Figura: Quantidade de artigos por técnica

Comparação da técnica proposta com as da literatura correlata

As principais vantagens da técnica proposta, em relação as da literatura correlata, são considerar as dependências de entrada e saída, semântica e a frequência de atividades.
Além disso, não há requesito de dados de proveniência, rede social, confiança entre usuários ou tipo de atividade: i) Shim; ii) Simples; e/ou iii) Subworkflow.

Solução proposta

A solução proposta neste mestrado recomenda atividades usando três conceitos importantes na área de *workflows* científicos: i) frequência de atividades; ii) compatibilidade entre entrada e saída; e ii) semântica de atividades

Desenvolvimento da Ontologia

A ontologia foi desenvolvida usando a metodologia *Skeletal*, que contém as seguintes fases:

1. Identificar a finalidade;
2. Construção da ontologia:
   1. Captura da ontologia;
   2. Codificação da ontologia;
   3. Integração com ontologias existentes;
3. Validação;
4. Documentação.

Matriz para técnicas da literatura

Tabela: Exemplo de matriz de entrada para técnicas da literatura correlata

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

Matriz para técnicas da classificação

São usadas as 59 atividades mais frequentes, para garantir o balanceamento do classificador estas são replicadas.

#	*Workflow*	Ativ 01	Ativ 02	. . .	Ativ 279	Ativ 280	Rótulo
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

Técnica proposta

Para explicar a técnica proposta será usada a figura:

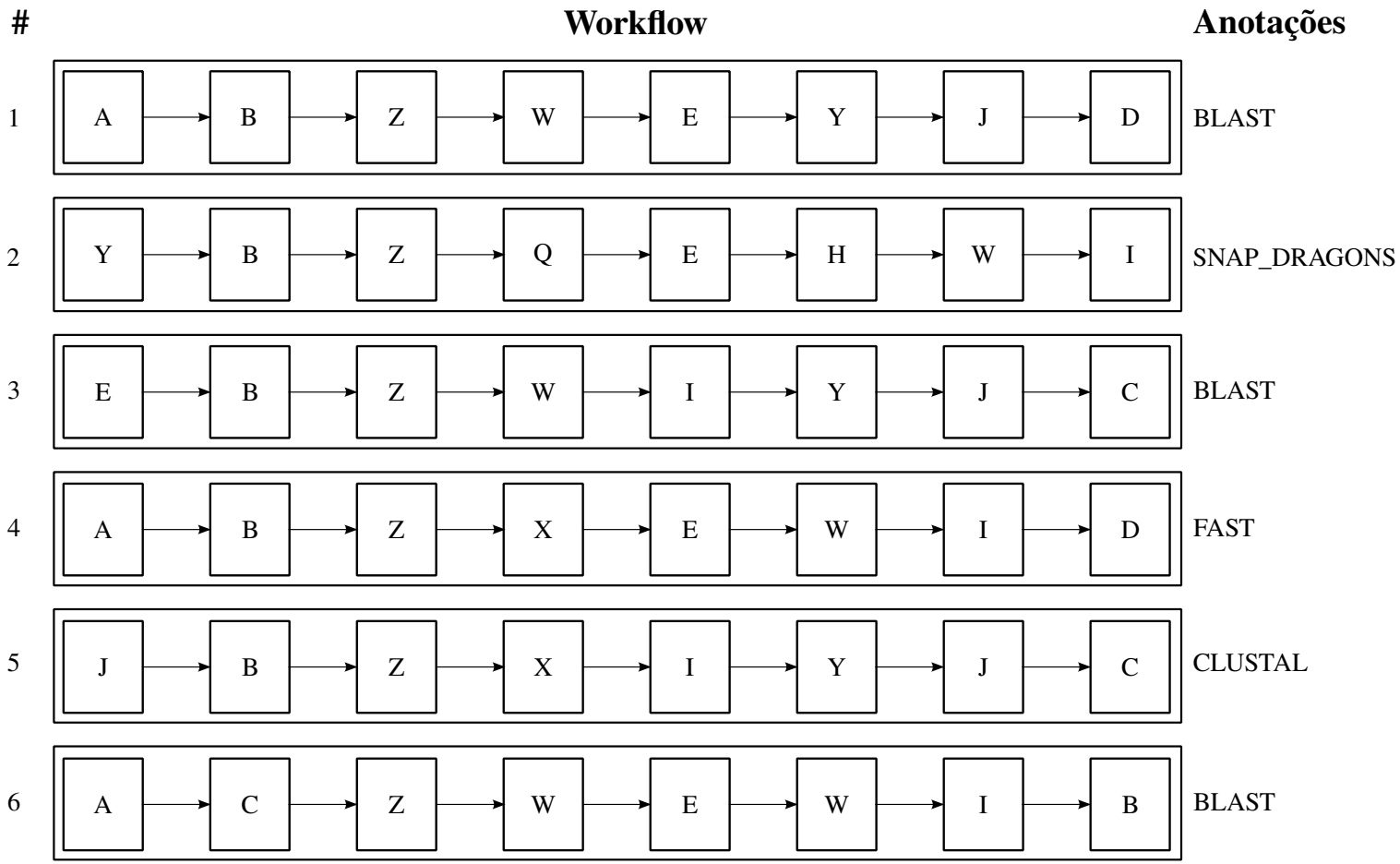

Figura: Exemplo de banco de dados de *workflows* científicos

Técnica proposta

Recomendação para a atividade Z ordenada por frequência e conceito ontológico.

Posição na Lista	Ativ	Frequência	Anotação Atividade
1	W	3	BLAST
2	X	2	FAST, CLUSTAL
3	Q	1	SNAP DRAGONS
⋮	⋮	⋮	⋮
280	⋮	⋮	⋮

Comparação dos experimentos

Resultados dos sistemas de recomendação.

#	**Técnica**	**S@1**	**S@5**	**S@10**	**S@50**	**S@100**	**S@280**	**MRR**
1	Aleatório	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	Naive $Bayes_C$	0,0274	0,1507	0,3425	0,6301	1,0000	1,0000	0,114
7	$Binomial_R$	0,0822	0,1918	0,2055	0,8493	1,0000	1,0000	0,136
8	Rede $neural_R$	0,1096	0,2603	0,2603	0,2603	1,0000	1,0000	0,154
9	$MARS_R$	0,1233	0,2055	0,2192	0,7260	1,0000	1,0000	0,167
10	SVM_R	0,1233	0,3151	0,4932	0,8493	1,0000	1,0000	0,238
11	$CART_R$	0,1370	0,1370	0,2603	0,6164	1,0000	1,0000	0,114
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

Comparação

1. O aumento de informação melhorou a recomendação.
2. Regressores foram melhores que classificadores (com exceção do SVM).
3. Classificadores compostos obtiveram um bom desempenho.
4. Converter valores contínuos com limiares possibilitou um bom desempenho no caso dos classificadores compostos.

Principais contribuições

1. Uma revisão sistemática sobre a área de recomendação de atividades em *workflows* científicos a qual poderá ser a base para trabalhos futuros.
2. Foi construída uma base de dados relacional de *workflows* científicos com suas respectivas atividades. Esta base será disponibilizada na íntegra para uso de outros trabalhos.
3. Foram implementadas diferentes técnicas da literatura correlata e foram comparados os resultados da recomendação dessas técnicas com os resultados da solução proposta.
4. Até o momento esta pesquisa de mestrado colaborou com a publicação de dois artigos cientfícos.

Considerações finais

Ao comparar todas as técnicas, foram constatados determinados aspectos do conjunto de dados, como o fato das atividades não serem independentes; o problema não ser linearmente separável; e que técnicas de agrupamento não se mostraram adequadas para solucionar este problema.
Com exceção do SVM, regressores apresentaram soluções mais precisas do que classificadores, além disso, adicionar informação nos sistemas de recomendação melhorou a precisão destes.

Trabalhos futuros

1. Usar outros classificadores compostos na recomendação de atividades;
2. Criar novas estratégias de recomendação baseadas em redes sociais de pesquisadores ou seus grupos de pesquisa;
3. Obter informação sobre proveniência de *workflows* e adicionar esta aos sistemas de recomendação;

Trabalhos futuros

1. Usar atividades de outros SGWC e/ou de outras áreas de pesquisa (além da bioinformática);
2. Estudar a relação entre a distribuição dos dados de entrada (atividade), sua esparsidade e a relação que ambas possuem com o aumento ou redução da precisão das recomendações;
3. Utilizar técnicas de redução de dimensionalidade para o conjunto de dados de entrada;
4. Adaptar o classificador SVM para considerar ontologias durante a maximização da margem ótima.

Publicações

1. Digiampietri, Luciano A. ; Perez-Alcazar, Jose J. ; Santiago, C. R. N. ; Oliveira, Guilherme A. ; Khouri, Adilson L. ; Araujo, Jonatas C. . A Framework for Automatic Composition of Scientific Experiments: Achievements, Lessons Learned and Challenges. VIII Brazilian e-Science Workshop (BreSci 2014), 2014, Brasília, Distrito Federal, Brasil. Anais do XXXIV Congresso da Sociedade Brasileira de Computação (CSBC 2014), 2014 (Publicado)
2. KHOURI, A. L. ; DIGIAMPIETRI, L. A. . A Systematic Review About Activities Recommendation in Workflows. In: 12ª Conferência Internacional sobre Sistemas de Informação e Gestão de Tecnologia, 2015, São Paulo. Anais da 12ª Conferência Internacional sobre Sistemas de Informação e Gestão de Tecnologia, 2015. v. 1. p. 14.(Publicado)

Obrigado

Obrigado por assistirem minha apresentação.

Agradecimentos

Agradecemos a Pró-Reitoria de Pós-Graduação da Universidade de São Paulo (USP) e a agência CAPES que forneceram bolsas de estudo para o estudante. Permitindo completar esse mestrado com publicações na área de computação. Além disso, agradecemos o professor Dr. Clodoaldo Aparecido de Lima por sanar dúvidas referentes a técnica SVM.