Desarrollo de técnicas para recomendar actividades en flujos de trabajo científicos: un enfoque basado en ontologías.

Alumno: Adilson Lopes Khouri

Orientador: Prof. Dr. Luciano Antonio Digiampietri

20 de octubre de 2019

Resumen

Introducción

1. *e-Science*.
2. Sistemas de gestión de flujos de trabajo científicos.
   - Ver grandes cantidades de datos.
   - Cálculos matemáticos.
   - Análisis del genoma.
3. Evitar escribir funciones/métodos existentes.
4. Gran cantidad de actividades.
5. Sistema para recomendar actividades.

Objetivo general

Este grado de maestria tiene como objetivo especificar e implementar una técnica de recomendación de actividad en flujos de trabajo científicos que combine:

1. Ontologías
2. Frecuencia de pares de actividades
3. Actividades de entrada y salida

Objetivos específicos

1. Construir una base de datos de flujos de trabajo científicos
2. Modelado de recomendaciones de actividad como un problema de clasificación/regresión
3. Comparación entre diferentes técnicas de literatura y soluciones propuestas

Conceptos Fundamentales

1. Sistemas de gestión de flujos de trabajo científicos.
2. Sistemas de recomendación.
3. Recomendación sobre flujos de trabajo científicos.
4. Ontologías.
5. Recomendación basada en bases de datos de flujos de trabajo científicos.
6. Validación de métricas.
7. Recomendación de la base de datos de flujo de trabajo.
8. Clasificadores y regresores.

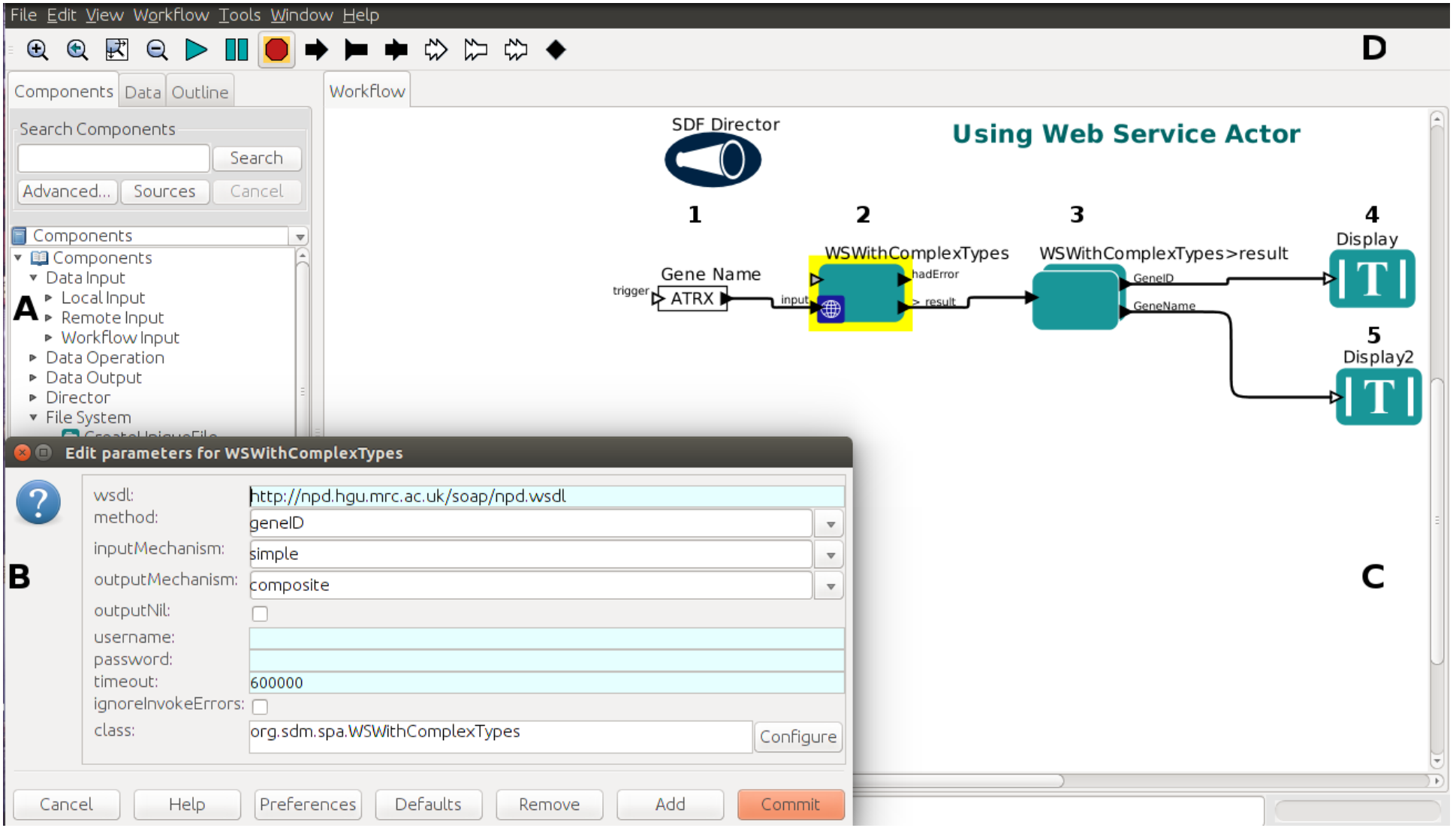


Figura: Ejemplo de sistema de gestión de flujo de trabajo científico.

Los sistemas de recomendación están destinados a **sugerir elementos útiles** a los usuarios:

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

La función de utilidad u no está definida para todo el espacio $C \times S$, esto obliga a los sistemas de recomendación a extrapolar el espacio conocido.

Algunas estrategias utilizadas en los sistemas de recomendación:

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

Recomendar actividades en flujos de trabajo científicos requiere, además de la extrapolación mencionada anteriormente, considerar las seguientes restricciones:

1. Dependencia entre entrada y salida de actividades
2. Dependencia semántica
3. El orden de las actividades

Ontología es un modelo para la representación del conocimiento, que se puede utilizar para anotar semánticamente las actividades.

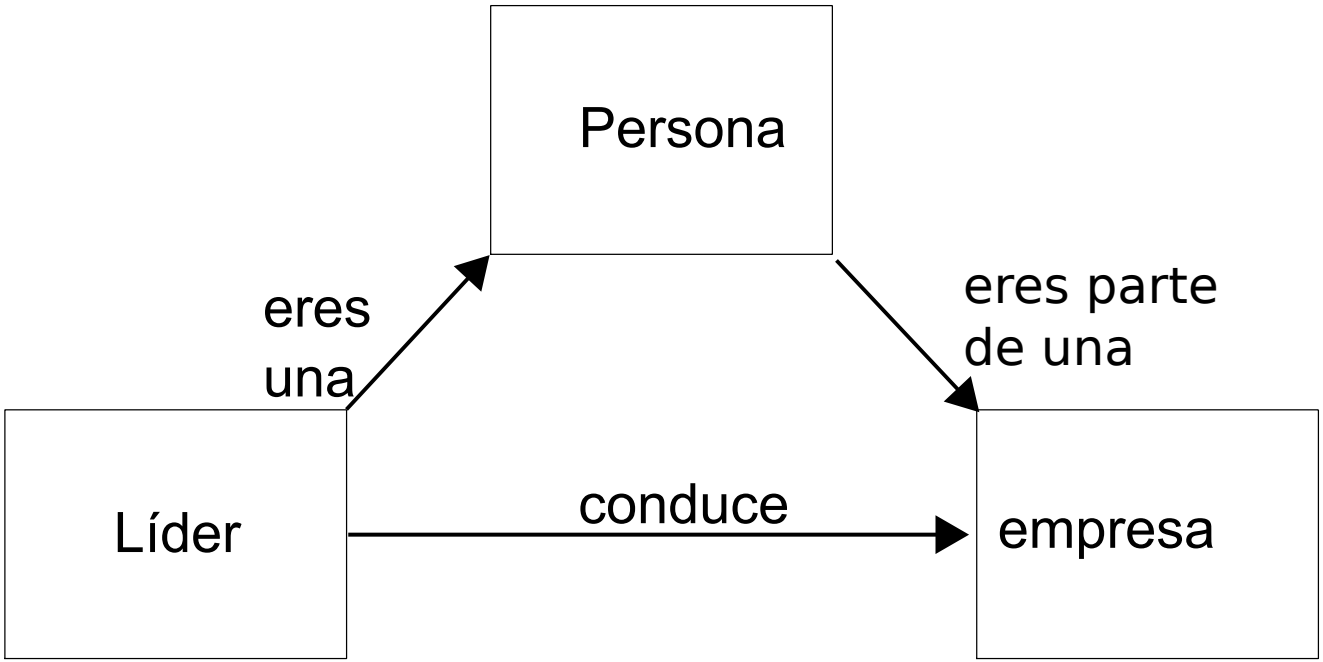


Figura: Ejemplo de ontología

Los experimentos serán testeados por *validación cruzada 10 veces*, cada ronda calculará las métricas:

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

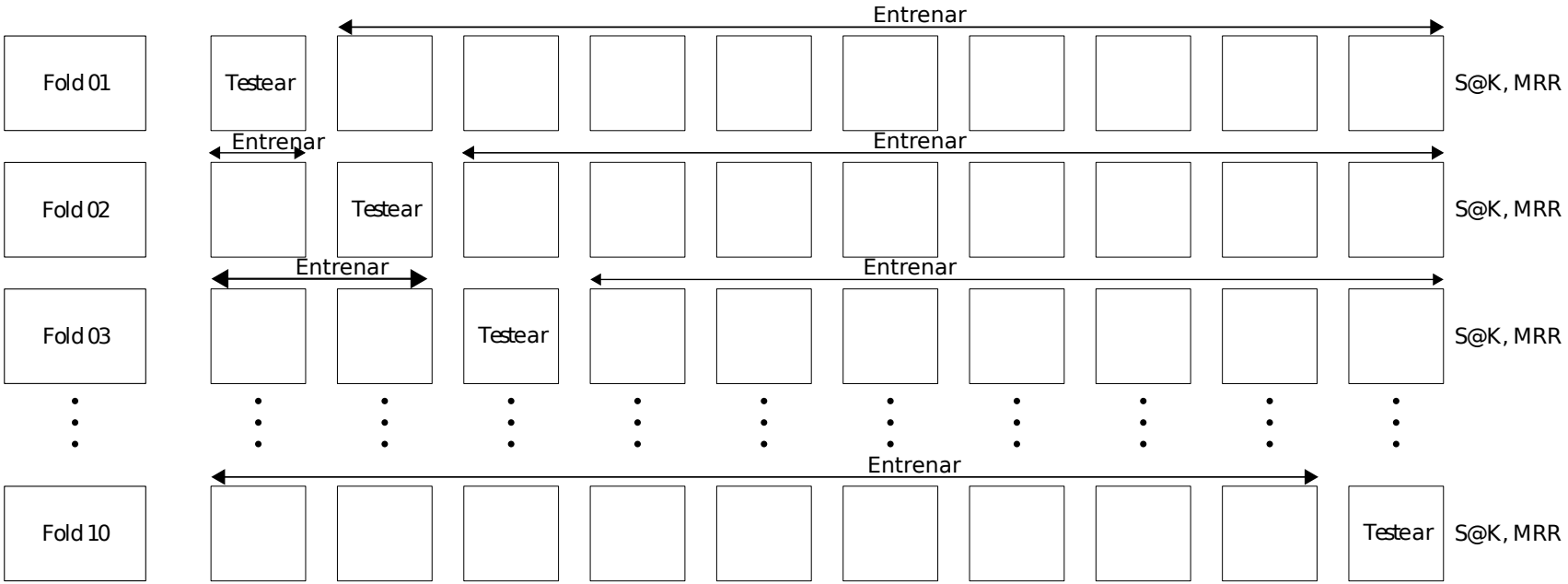


Figura: Ejemplo de *10-fold cross validation*

Recomendación de base de datos de flujo de trabajo

1. Frecuencia
2. *itemsets*.

Recomendación de clasificadores

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

Recomendación de los regresores

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

Recomendación de clasificadores compuestos

1. SVM;
2. Rotation Forest.

Revisión sistemática

La revisión de la literatura empezó con un estudio exploratorio seguido de una revisión sistemática. Por lo tanto, fue posible:

1. Encontrar un área de recomendación de vanguardia para flujos de trabajo científicos.
2. Comprender el problema.
3. Encuentrar términos específicos del área.
4. Definir palabras clave

Conducir

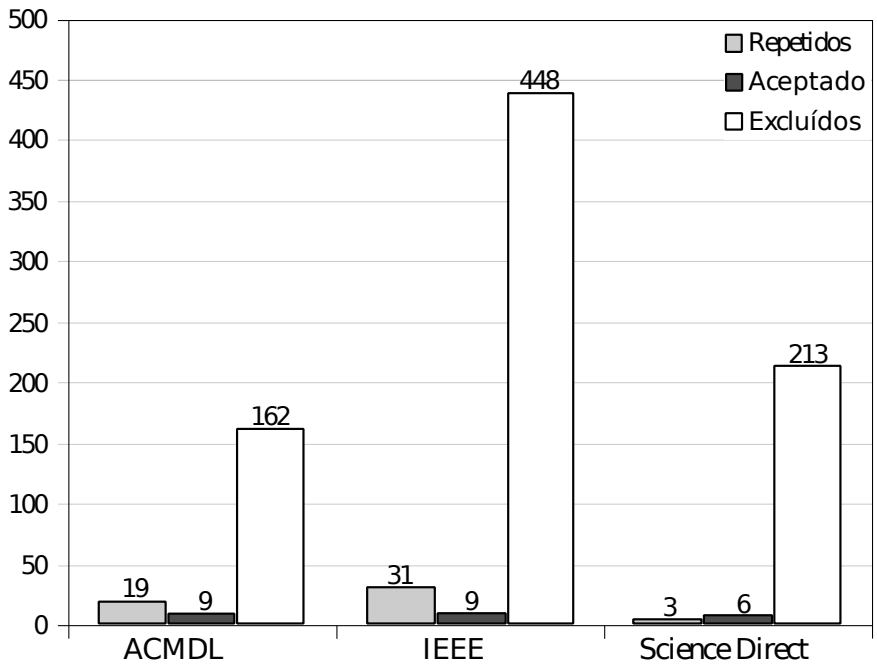


Figura: Número de artículos por técnica.

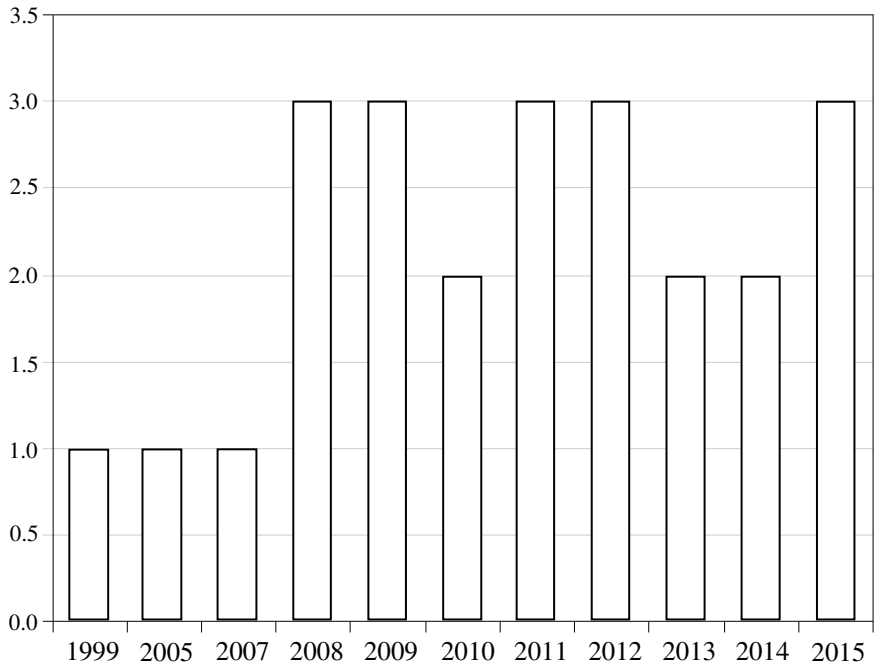
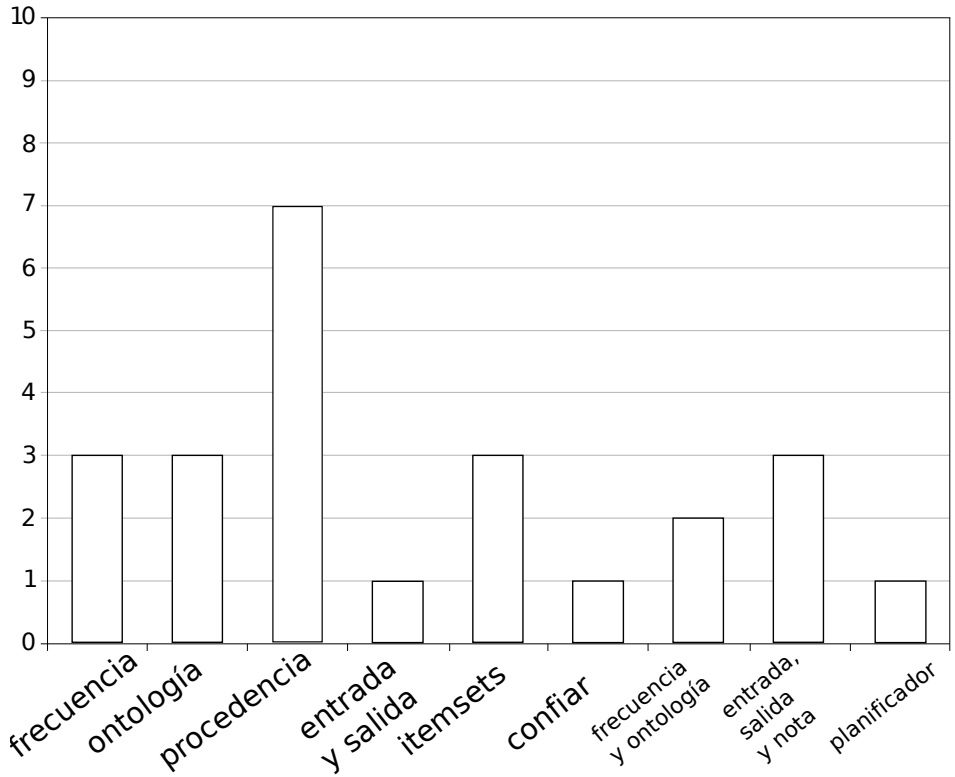


Figura: Artículos por año de publicación

Ejecución

Se observa que la técnica de procedencia es la más utilizada seguida de: i) Frecuencia; ii) entrada y salida; iii) *itemsets*; y iv) ontologías.



Comparación de la técnica propuesta con la literatura relacionada.

Las principales ventajas de la técnica propuesta, en relación con las de la literatura relacionada, son considerar las dependencias de entrada y salida, la semántica y la frecuencia de las actividades.
Además, no se requieren datos de procedencia, redes sociales, recomendacion por confianza entre usuarios o tipo de actividad: i) Shim; ii) Simples; I/O iii) Subflujo de trabajo.

Solución propuesta

La solución propuesta en este Máster recomienda actividades que utilizan tres conceptos importantes en el área de los flujos de trabajo científicos: i) frecuencia de las actividades; ii) compatibilidad entre entrada y salida; y ii) semántica de actividad

Desarrollo de ontologías

La ontología se desarrolló utilizando la metodología *Skeletal*, que contiene las siguientes fases:

1. Identificar el propósito
2. Construcción de ontologías:
   1. Captura de ontologías
   2. Codificación de ontología;
   3. Integración con ontologías existentes
3. Validación
4. Documentación

Matriz de técnicas de literatura

Cuadro: Ejemplo de matriz de entrada para técnicas de literatura relacionadas

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

Matriz para técnicas de clasificación

Las actividades más comunes que se utilizan para garantizar el equilibrio del clasificador en el que se replican, son 59

#	***Workflow***	**Ativ 01**	**Activ 02**	. . .	**Activ 279**	**Activ 280**	**Eti**
1	01	1	0	. . .	0	0	
2	01	1	0	. . .	0	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
59	01	1	0	. . .	0	0	
1	01	0 (eliminado)	1 (añadido)	. . .	1	0	
2	01	0 (eliminado)	0	. . .	1 (añadido)	0	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
59	01	0 (eliminado)	0	. . .	0	1 (añadido)	
	⋮						
1	73	1	1	. . .	0	0	
2	73	1	1	. . .	0	0	

Técnica propuesta

Para explicar la técnica propuesta se utilizará la figura:

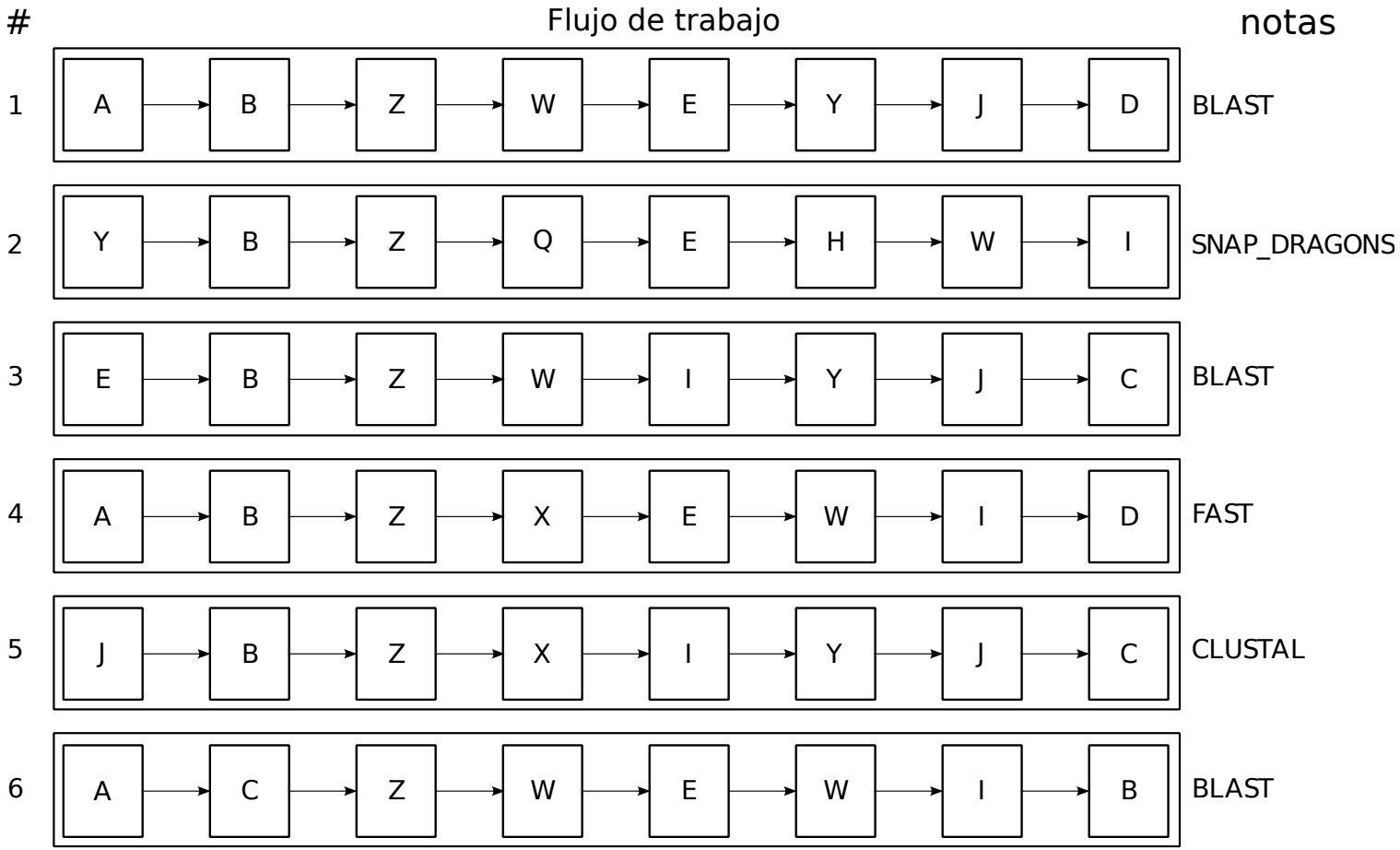


Figura: Ejemplo de base de datos de flujos de trabajo científicos

Técnica propuesta

Recomendación para la actividad Z ordenada por frecuencia y concepto ontológico.

Posición de la lista	Activ	Frecuencia	Actividad de notas
1	W	3	BLAST
2	X	2	FAST, CLUSTAL
3	Q	1	SNAP DRAGONS
⋮	⋮	⋮	⋮
280	⋮	⋮	⋮

Comparación de experimentos

Resultados de los sistemas de recomendación.

#	**Técnica**	**S@1**	**S@5**	**S@10**	**S@50**	**S@100**	**S@280**	**MRR**
1	Random	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 neuronal$_C$	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 neuronal$_R$	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

Comparación

1. El aumento de la información ha mejorado la recomendación.
2. Los regresores fueron mejores que los clasificadores (excepto SVM).
3. Los clasificadores compuestos funcionaron bien.
4. La conversión de valores continuos con limites proporcionó un buen rendimiento para los clasificadores compuestos.

Mayores contribuciones

1. Una revisión sistemática del área recomendada de actividades en los flujos de trabajo científicos que puede ser la base para el trabajo futuro.
2. Se construyó una base de datos relacional de flujos de trabajo científicos con sus respectivas actividades. Esta base estará disponible en su totalidad para su uso por otras obras.
3. Se implementaron diferentes técnicas de la literatura relacionada y los resultados de la recomendación de estas técnicas se compararon con los resultados de la solución propuesta.
4. Hasta ahora, la investigación de este maestro ha contribuido a la publicación de dos artículos científicos.

Consideraciones finales

Al comparar todas las técnicas, se encontraron ciertos aspectos del conjunto de datos, como el hecho de que las actividades no eran independientes; el problema no es linealmente separable; y que las técnicas de agrupamiento no eran adecuadas para resolver este problema.

Con la excepción de SVM, los regresores tienen soluciones más precisas que los clasificadores, y también agregan información a los sistemas de recomendación mejorarron su precisión

Trabajos futuros

1. Usar otros clasificadores compuestos al recomendar actividades
2. Crear nuevas estrategias de recomendación basadas en las redes sociales de investigadores o sus grupos de investigación
3. Obtener información sobre la procedencia de flujos de trabajo y agréguela a los sistemas de recomendación;

Trabajos futuros

1. Utilizar actividades de otras áreas de investigación de SGWC y / u otras (además de bioinformática)
2. Para estudiar la relación entre la distribución de datos de entrada (actividad), su escasez y la relación que ambos tienen con el aumento o reducción de la precisión de las recomendaciones
3. Utilizar técnicas de reducción de dimensionalidad para el conjunto de datos de entrada;
4. Adaptar el clasificador SVM para considerar las ontologías durante la maximización de la margen óptima.

Publicaciones

1. KHOURI, ADILSON LOPES; DIGIAMPIETRI, LUCIANO ANTONIO . Combining Artificial Intelligence, Ontology, and Frequency-based Approaches to Recommend Activities in Scientific Workflows. REVISTA DE INFORMÁTICA TEÓRICA E APLICADA: RITA, v. 25, p. 39-47, 2018. (Publicado)
2. 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)

Publicaciones

1. 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)

Gracias

Gracias por ver mi presentación.

Agradecimientos

Agradecemos al Decano de Estudios de Posgrado de la Universidad de São Paulo (USP) y a la agencia CAPES que brindó becas para el estudiante y a profesor Roberto Ortiz de “Casa de Idiomas y Cultura” que ayudou con la traduccion. Permitiendo completar este máster con publicaciones en el área de informática. Además, agradecemos al profesor Dr. Clodoaldo Aparecido de Lima por responder preguntas sobre la técnica SVM.