Resumen

La ciencia de datos es esencial para extraer valor de los datos. Sin embargo, la parte más tediosa del proceso, la limpieza y manipulación de datos, requiere una serie formateos y acciones mayormente manuales. Este proceso todavía se resiste a la automatización en parte porque el problema depende en gran medida de la información del dominio, que se convierte en un cuello de botella para los sistemas más modernos a medida que aumenta la diversidad de dominios, formatos y estructuras de los datos.

En esta tesis nos enfocamos en generar algoritmos que aprovechen el conocimiento del dominio para la automatización de partes del proceso de limpieza y manipulación de datos. Demostramos cómo las técnicas generales de inducción de programas, en lugar de lenguajes específicos de dominio, se pueden aplicar de manera flexible a problemas donde el conocimiento es importante, mediante el uso dinámico del conocimiento específico del dominio. De manera más general, sostenemos que el éxito en la automatización reside en una combinación de enfoques de aprendizaje basados en conocimiento y dinámicos. Proponemos varias estrategias para seleccionar o construir automáticamente el conocimiento previo apropiado para varios escenarios de manipulación de datos. La idea clave se basa en elegir las mejores primitivas especializadas de acuerdo con el contexto del problema a resolver.

Para esto abordamos dos escenarios. En el primero, manejamos datos personales (nombres, fechas, teléfonos, etc.) que se presentan en formatos de texto muy diferentes y deben ser transformados a un formato unificado. El problema radica en cómo construir una transformación compositiva a partir de un gran conjunto de primitivas en el dominio (por ejemplo, manejar meses, años, días de la semana, etc.). Para ello, desarrollamos un sistema (BK-ADAPT) que guía la búsqueda a través del conocimiento previo extrayendo varias meta-características de los ejemplos que caracterizan el dominio de la columna de datos. En el segundo escenario nos enfrentamos a la transformación de matrices de datos en lenguajes de programación genéricos como R, utilizando una matriz de entrada y algunas celdas de la matriz de salida como ejemplos. También desarrollamos un sistema guiado por una búsqueda basada en árboles (AUTOMAT[R]IX) que incorpora varias restricciones, probabilidades previas para las primitivas y sugerencias textuales para aprender eficientemente las transformaciones.

Con estos sistemas, mostramos que la combinación de programación inductiva con la selección dinámica de las primitivas apropiadas a partir del conocimiento previo es capaz de mejorar los resultados de otros enfoques de manipulación de datos actuales y más específicos.