Evaluación de la Calidad de los Datos en Sistemas de Integración de Datos

La necesidad de acceder de manera uniforme a múltiples fuentes de datos es cada día más fuerte y generalizada, especialmente en el contexto de aplicaciones para toma de decisiones, las cuales necesitan de un análisis comprensivo y exploratorio de los datos. Con el desarrollo de Sistemas de Integración de Datos (SID), la calidad de la información se ha transformado en una propiedad de primer nivel, cada vez más requerida por los usuarios.

Esta tesis trata sobre la evaluación de la calidad de los datos en los SID. En particular, se abordan los problemas de la evaluación de la calidad de los datos que responden a consultas de usuarios y la satisfacción de las exigencias de dichos usuarios en términos de calidad. Se analiza también la utilización de medidas de calidad para mejorar el diseño del SID e incrementar la calidad de los datos. Nuestro enfoque consiste en estudiar un factor de calidad a la vez, analizando su impacto en el SID, proponiendo técnicas para su evaluación y proponiendo acciones para su mejora. Entre los factores de calidad que se han propuesto en la literatura, esta tesis analiza dos de los más usados: *la frescura* y *la exactitud* de los datos.

Analizamos las diferentes definiciones y métricas que se han propuesto para la frescura y la exactitud de los datos y abstraemos las propiedades del SID que juegan un rol importante en su evaluación. El análisis de cada factor se resume en una taxonomía, la cual permite comparar los trabajos existentes y resaltar los problemas abiertos.

Proponemos un marco de trabajo que modela los diferentes elementos relacionados a la evaluación de la calidad: fuentes de datos, consultas de usuarios, procesos de integración del SID, propiedades del SID, medidas de calidad y algoritmos de evaluación de la calidad. En particular, los procesos de integración del SID se modelan como flujos de trabajo, cuyas actividades realizan las tareas de extracción, integración y entrega de los datos a los usuarios. Nuestro soporte de razonamiento para la evaluación de la calidad es un grafo acíclico dirigido, llamado grafo de calidad, que tiene la misma estructura del SID y está etiquetado con las propiedades del SID que son relevantes para la evaluación de la calidad. Los algoritmos de evaluación de la calidad toman como entrada los valores de calidad de los datos fuentes y las propiedades del SID y combinan dichos valores obteniendo una medida de la calidad de los datos retornados por el SID. Los algoritmos se basan en la representación de grafo y combinan los valores de las propiedades mientras recorren el grafo. Los mismos pueden instanciarse para tener en cuenta las propiedades que influencian la calidad en una aplicación concreta. La idea detrás del marco de trabajo es definir un contexto flexible que permita la especialización de los algoritmos para escenarios de aplicación específicos.

Los valores de calidad obtenidos durante la evaluación se comparan con los valores exigidos por los usuarios. Si las exigencias de calidad no son satisfechas, se pueden realizar acciones de mejora al SID. Sugerimos un conjunto básico de acciones de mejora que pueden componerse para mejorar la calidad en SID concretos. Para mejorar la frescura de los datos proponemos analizar el SID a diferentes niveles de abstracción, de manera de identificar sus puntos críticos (las porciones del SID que causan la no satisfacción de las exigencias de frescura) y concentrar la aplicación de acciones de mejora sobre esos puntos. Para mejorar la exactitud de los datos proponemos partir los resultados de las consultas en áreas (algunos atributos, algunas tuplas) de exactitud homogénea. Esto permite que las aplicaciones de los usuarios desplieguen solamente los datos más exactos, filtren los datos que no satisfacen las exigencias de exactitud o desplieguen los datos en camadas según sus exactitudes. Nuestro enfoque se diferencia de los enfoques existentes de selección de fuentes porque podemos seleccionar áreas de buena exactitud en lugar de sólo seleccionar fuentes enteras.

Las principales contribuciones de esta tesis son: (i) un análisis detallado de los factores de calidad frescura y exactitud, (ii) la propuesta de técnicas y algoritmos de evaluación y mejora de la frescura y la exactitud de los datos, y (iii) un prototipo de evaluación de la calidad utilizable en contextos prácticos de diseño de SID.