Proyecto: Análisis de Factores de Calidad en Sistemas de Información Multi-fuentes
Informe Técnico

A) Objetivos generales y específicos.

Objetivo general:

Proponer un marco de trabajo (framework) para el manejo de propiedades de calidad en un Sistema de Información Multi-fuentes (en adelante, SIMF), que nos permita evaluar la calidad del sistema, tomar decisiones de diseño del mismo, y resolver los principales problemas relacionados con los cambios en la calidad de las fuentes de datos.

Objetivos específicos:

- Identificar un conjunto minimal de propiedades de calidad relevantes para un SIMF, que serán tomadas como base para el proyecto. No se trata de identificar un conjunto completo de propiedades, sino un conjunto mínimo que pueda usarse como base representativa para la investigación.
- Analizar técnicas de evaluación para las propiedades de calidad seleccionadas y determinar las características de las fuentes o del sistema que influyen en la calidad del SIMF.
- Analizar el impacto de las propiedades de calidad de las fuentes y requerimientos de calidad del sistema, en el diseño del SIMF, considerando solamente las propiedades de calidad seleccionadas.
- Estudiar el problema de gestión de cambios en los valores de calidad de las fuentes, para las propiedades de calidad seleccionadas, y proponer estrategias para el manejo.
- Especificar un mecanismo general para evaluación de la calidad en un SIMF.
- Especificar técnicas generales para el manejo de cambios de calidad en las fuentes en un SIMF.
- Implementar un prototipo de un marco de trabajo para el manejo de propiedades de calidad en un SIMF.

B) Actividades

Se ha trabajado en base a las siguientes actividades:

- Tesis de doctorado de Verónika Peralta
  En esta tesis se ha trabajado en el estudio de los factores de calidad “frescura” y “exactitud” (“freshness y “accuracy”), en la definición de un marco de trabajo donde evaluar estos factores en un SIMF, y en los algoritmos de evaluación de dichos factores para distintos escenarios posibles. Se generaron reportes técnicos [Per06-1] [Per06-2], y se publicaron artículos [PB05], [GPB05], [KPSX05].
  La tesis fue finalizada y fue defendida con éxito en noviembre de 2006 [Per06-3].

- Tesis de doctorado de Adriana Marotta
  En ella se ha trabajado en el tema de cambios en los factores de calidad del SIMF. El estudio se ha centrado en el manejo de los factores "frescura" y "exactitud", retroalimentándose con los resultados de la tesis de Verónika Peralta. Se generó un reporte técnico [MR05-1] y se publicaron 2 artículos [MR05-2] [Mar06]. Además se sometió un artículo con los últimos avances de la tesis, del cual todavía no se sabe si es aceptado o no [MR07]. Esta tesis tiene un avance del 80 %.

- Trabajo dirigido: Evaluación de calidad en el Data Warehouse de Enseñanza de la Facultad de Ingeniería.
  El objetivo de este trabajo es aplicar las propuestas de evaluación de calidad al DW de Enseñanza de la Facultad de Ingeniería, el cual es un caso real de sistema multifuente.
  El trabajo fue llevado adelante por participantes que cumplieron uno de los siguientes dos roles: el rol de tutor y el rol de estudiante. Constó de tres módulos que correspondieron a las tres etapas necesarias para realizar la evaluación de calidad: (1) Estudio del proceso de transformación de datos desde las fuentes hacia el DW, (2) Medición de la Calidad de los datos fuentes, y (3) Estimación de la calidad en el DW. En las distintas etapas algunos participantes fueron cambiando.
  Las tutoras daban las pautas y apoyaban el desarrollo del trabajo realizado por los estudiantes. Este trabajo implicó, además del desarrollo en sí mismo, reuniones semanales de las personas participantes, donde se discutían los distintos problemas a resolver y las tutoras marcaban el rumbo del trabajo. Se han generado dos reportes técnicos [ETG06] y [ETM07], que corresponden a las primeras dos etapas del trabajo, y un documento interno del proyecto acerca de la tercera etapa.

- Implementación de prototipo de herramienta.
  Se trabajó en la implementación de una herramienta para evaluación y manejo de cambios de calidad de un SIMF. Este trabajo se hizo en torno a un Proyecto de Grado de la carrera Ing. en Computación, tutoreado por integrantes del proyecto. Los resultados del proyecto de grado se encuentran en [RS06].

- Trabajo conjunto con Profesor visitante
  Se realizó un trabajo conjunto con el Profesor Alberto Abelló, de la Universidad Politécnica de Cataluña, España. En el mismo se aplicaron conceptos de Calidad en SIMF al caso particular de Sistemas OLAP. A partir de este trabajo se generó una publicación en una conferencia internacional [MPA06].

- Interacción con investigadores del Area Investigación Operativa
  Se realizaron reuniones con investigadores del grupo IO del Instituto de Computación, con el objetivo de validar los aspectos de nuestro trabajo que involucran aplicación de Técnicas Probabilísticas y recibir sus valiosos aportes. Este tipo de técnicas las estamos aplicando en algunas estrategias que proponemos para el manejo de cambios de calidad.

C) Realización de lo planteado en el proyecto

En grandes rasgos, las fases planteadas en el proyecto eran:

- estudio de un conjunto pequeño de propiedades de calidad
- propuesta de un marco de trabajo y técnicas para la evaluación de dichas propiedades en SIMF
- impacto de las propiedades de calidad consideradas, en el diseño del SIMF
- propuesta de técnicas para manejo de cambios en las propiedades de calidad, basadas en modelos probabilísticos
- implementación de un prototipo de herramienta para evaluación de calidad en SIMF

- generalización de las propuestas para evaluación y manejo de cambios, a un conjunto amplio de propiedades de calidad

Todas estas fases fueron cumplidas completamente, excepto la última. Esta última fase fue sustituida por un trabajo de experimentación de las técnicas propuestas en casos reales. El cambio se debió a que al avanzar la investigación se hizo fundamental, como forma de validación de las propuestas, realizar dicha experimentación, aunque en el proyecto original no había sido considerado. Por otro lado, al profundizar en el estudio de las propiedades de calidad y en las propuestas para el manejo de éstas, nos encontramos con que cada propiedad de calidad tiene características muy particulares y diferentes entre sí. Esto hace que las soluciones para cada una de ellas sean muy específicas. De todas formas se pudo intuir que las ideas generales de las soluciones pueden ser aplicadas a cualquier otra propiedad de calidad.

Se realizaron dos experimentaciones: (1) aplicación de las técnicas de evaluación de calidad en el Data Warehouse de Enseñanza de la Facultad de Ingeniería, y (2) aplicación de propuestas de manejo de cambios de calidad en un SIMF con datos de películas de cine y opiniones de espectadores, construido a partir de fuentes de datos de la Web. Estas experimentaciones se realizaron en su mayor parte durante el último período (2º. año) del desarrollo del proyecto.

D) Principales resultados obtenidos

- Definición y estudio en profundidad de dos propiedades de calidad: “frescura” y “exactitud” [Per06-2]. Se estudiaron distintos escenarios posibles y métricas para estas propiedades. [GPB05], [KPSX05] [Per06-1]
- Propuesta de un marco de trabajo para el manejo de la calidad. Se especificó un marco de trabajo en donde se puede modelar el grafo de transformación de datos del sistema multifuente, con sus propiedades que definen distintos escenarios. En este marco de trabajo se pueden realizar los cálculos necesarios para la evaluación de calidad y para el manejo de cambios de la calidad de las fuentes. [Per06-3]
- Propuesta de técnicas de evaluación de calidad en el SIMF para las propiedades “frescura” y “exactitud”. Se propusieron los algoritmos de cálculo de los valores de calidad del sistema a partir de los valores de calidad de las fuentes, para estas propiedades. [Per06-3]
- Caracterización del problema de cambios en la calidad del SIMF. Se estudiaron y definieron las características particulares del problema de cambios, realizándose una comparación con el problema, ya conocido, de evolución de esquemas fuentes en un sistema de integración de esquemas. [MR05-1]
- Propuesta para manejo de cambios en la calidad de un SIMF. Se propone un mecanismo que se basa principalmente en tres técnicas: (1) modelado probabilístico del comportamiento de la calidad de cada fuente y del sistema, (2) manejo de eventos a través de reglas para detectar cambios relevantes, y (3) determinación de acciones de corrección del sistema en los casos en que ocurrió un cambio relevante. Esta propuesta está siendo documentada en el trabajo de tesis de Adriana Marotta, a concluirse en el correr del presente año. Resultados intermedios se encuentran en [Mar06] [MR07].
- Prototipo del marco de trabajo para manejo de calidad. Herramienta donde se modela el sistema multifuente y sus propiedades de calidad, y se pueden definir y ejecutar algoritmos de evaluación de la calidad del sistema a partir de la calidad de las fuentes. También ofrece funcionalidades útiles para el manejo de cambios de calidad. [RS06]
- Implementación de una medición de calidad en un caso real. [ETM07]

Conclusiones:

Se propone un marco de trabajo para el manejo de propiedades de calidad en SIMF, el cual es además implementado en un prototipo. Para las propiedades de calidad "frescura" y "exactitud" se proponen técnicas concretas para:

- cálculo de la calidad en el SIMF a partir de los valores en las fuentes

- manejo de los cambios en los valores de calidad del sistema

Se prueban algunas de las propuestas en dos casos reales, mostrando la viabilidad de la aplicación de las mismas.

E) Autoevaluación

La ejecución del proyecto ha aportado conocimientos originales, ya que tanto el marco de trabajo para manejo de calidad en sistemas multifuentes, como las soluciones específicas para calcular la frescura y la exactitud en dichos sistemas y también para manejar los cambios de estas propiedades de calidad, son propuestas completamente innovadoras. Estos problemas han sido muy escasamente abordados previamente por otros investigadores, y en esos casos se ha trabajado con enfoques diferentes al planteado en este proyecto. Los resultados obtenidos fueron publicados y presentados en diversos y prestigiosos eventos internacionales sobre Sistemas de Información.

Es también relevante el hecho de haberse implementado un prototipo de herramienta para el manejo de calidad (marco de trabajo), y de haberse experimentado la aplicación de las técnicas de evaluación de calidad y de manejo de cambios en casos reales.

La ejecución del proyecto sirvió como contexto de trabajo y fortaleció el desarrollo de dos tesis de doctorado. Por otra parte generó trabajos de posgrado para estudiantes que se encuentran en el comienzo de sus estudios de Maestría. Finalmente, dio contexto al desarrollo de un proyecto de grado de fin de la carrera de Ingeniero en Computación.

Por lo expresado anteriormente se considera que el desarrollo del proyecto y los resultados obtenidos son ampliamente satisfactorios.

F) Publicaciones y Materiales de Difusión **(se adjuntan documentos)**

[ETG06] L. Etcheverry, P. Gatto, S. Tercia. *Análisis del proceso de carga del Sistema de Data Warehousing de Enseñanza de la Facultad de Ingeniería*. Reporte Técnico INCO RT 06-06. ISSN 0797-6410. InCo, Facultad de Ingeniería, Universidad de la República. Montevideo, Uruguay. Abril 2006.

[ETM07] L. Etcheverry, S. Tercia, A. Marotta, V. Peralta. *Medición de la Exactitud de Datos en Sistemas Fuentes: Un Caso de Estudio.* Reporte Técnico INCO (en proceso de numeración). InCo, Facultad de Ingeniería, Universidad de la República. Montevideo, Uruguay. Abril 2006.

[GPB05] D. Grigori, V. Peralta, M. Bouzeghoub. *Service Retrieval Based on Behavioral Specifications and Quality Requirements*. 3rd International Conference on Business Process Management (BPM'2005). Nancy, France, Septembre 2005.

[KPSX05] D. Kostadinov, V. Peralta, A. Soukane, X. Xue. *Intégration de données hétérogènes basée sur la qualité*. Jornadas INFORSID'2005, Grenoble, France, Mayo 2005.

[Mar06] A. Marotta. *Managing source quality Changes in a Data Integration System.* Doctoral Consortium of 18th. Conference on Advanced Information Systems Engineering (CAISE'06). Luxembourg, Luxembourg, June, 2006.

[MPA06] A. Marotta, F. Piedrabuena, A. Abelló. *Managing Quality Properties in a ROLAP Environment.* Accepted in 18th. Conference on Advanced Information Systems Engineering (CAISE'06). Luxembourg, Luxembourg, June, 2006.

[MR05-1] A. Marotta, R. Ruggia. *Manejo de Cambios en la Calidad de las Fuentes en Sistemas de Integración de Datos.* Reporte Técnico INCO RT 05-10. ISSN 0797-6410. InCo, Facultad de Ingeniería, Universidad de la República. Montevideo, Uruguay. Setiembre. 2005.

[MR05-2] A. Marotta, R. Ruggia. *Managing Source Quality Changes in Data Integration Systems*. Second International Workshop on Data and Information Quality (DIQ'05) (in conjunction with CAISE). Porto, Portugal, June, 2005.

[MR07] A. Marotta, R. Ruggia. *Quality Changes Management in Data Integration Systems. A Probability Based Approach.* Submitted to 2nd. International Conference on Software and Data Technologies (ICSoft'07). Barcelona, Spain, July, 2007.

[PB05] V. Peralta, M. Bouzeghoub. *Data Freshness Evaluation in Different Application Scenarios.* Revue des Nouvelles Technologies de l'Information (RNTI), Vol E-5 (Extraction des connaissances : Etat et perspectives). ISBN 2.85428.707.x, 2006.

[Per06-1] V. Peralta. *Evaluating Data Freshness in web warehousing applications: A case of study*. Reporte interno, InCo, Universidad de la República, URUGUAY, Febrero 2006.

[Per06-2] V. Peralta. *Data Freshness and Data Accuracy – State of the Art.* Reporte Técnico, TR06-13, InCo, Universidad de la República, Montevideo, URUGUAY, Marzo 2006.

[Per06-3] Verónika Peralta. *Data Quality Evaluation in Data Integration Systems*. Tesis de Doctorado, Universidad de Versalles, FRANCIA – Universidad de la República, URUGUAY, Noviembre 2006.

[RS06] M. Ramos, R. Séttimo. *Herramienta para evaluación y configuración de la calidad en sistemas de información multi-fuente.* Proyecto de grado de la carrera Ingeniería en Comnputación, Facultad de Ingeniería, Universidad de la República. Tutoras: A. Marotta, V. Peralta. Mayo 2006. http://www.fing.edu.uy/~pgcaldat/

F) Trabajo Futuro

La investigación desarrollada en el marco de este proyecto va a continuarse, a corto plazo con la culminación de la tesis de doctorado sobre el manejo de cambios en la calidad de SIMF (de A. Marotta), y más a largo plazo, se continuará trabajando en el tema de calidad, extendiéndose al estudio de la calidad en otros contextos. Por ejemplo, en nuestro grupo (CSI, Instituto de Computación), se está empezando a trabajar en la *calidad de artículos* (documentos) por un lado, y por otro lado en la *calidad del diseño* de sistemas de información.

En cuanto a las futuras aplicaciones, se pretende continuar aplicando las propuestas en distintos dominios de aplicación. En particular, se está comenzando a trabajar con el área biológica; en el marco de un trabajo conjunto con el Instituto Pasteur, se está investigando como evaluar la calidad de información proveniente de experimentos biológicos.