ISSN: 1692-7257 Volumen 1 Número 5 año - 2005
Revista Colombiana de Tecnologías de Avanzada

QUALITY-DRIVEN IN QUERY PLANNING FOR WEB INFORMATION TECHNOLOGIES

EL MANEJO DE CALIDAD EN LA PLANIFICACIÓN DE CONSULTAS PARA TECNOLOGÍAS DE INFORMACIÓN WEB

[1]Ing. Bell Manrique L. [2]MSc. Jaime Alberto Guzmán L., [3]MSc. Francisco Javier M.

Universidad Nacional de Colombia Sede Medellín
Escuela de Sistemas, Facultad de Minas, Núcleo Robledo, Medellín
[1]Estudiante Maestría en Ingeniería de Sistemas, bmanriq@unalmed.edu.co
jaguzman@unalmed.edu.co, fjmoreno@unalmed.edu.co

Abstract: In the following article we present an overview of recent works oriented to the query planning with quality control in the Information Systems Based on Mediators. In our study the quality in plans and resulting information of the queries are important aspects. We describe proposals oriented to face the problem from a Web Information Systems viewpoint, by means of traditional query optimization, Multiagent Systems and Mediators Systems. Finally, we present some open problems in this field and we describe an initial proposal to face them.

Resumen: En el siguiente artículo se presenta una revisión de los recientes trabajos orientados a la Planificación de Consultas con control de calidad en los Sistemas de Información Basados en Mediadores, teniendo en cuenta la calidad tanto de los planes, como de la información resultado de las consultas. Se describen las propuestas orientadas a enfrentar el problema desde el marco de acción de los Sistemas de Información Web, pasando por la optimización de consultas tradicional, los Sistemas Multiagente y los Sistemas Mediadores. Por último se presentan algunos problemas abiertos en este tópico de investigación y se plantea una propuesta inicial para enfrentarlos.

Keywords: Query Planning, Information Mediators, Information Integration, Information Quality, Web Technologies.

1. INTRODUCCIÓN

El acceso integrado a información que reposa sobre múltiples fuentes de información heterogénea y distribuida disímiles en sintaxis, estructura y semántica, como es el caso específico de la Web es un problema importante en muchos dominios actuales.

Para manejar la integración de información se han creado varios componentes que interactúan entre sí, ofreciendo acceso integrado a datos desde un dominio específico, éstos son los *Mediadores* y los *Wrappers*.

Los *Mediadores* son aquellos componentes que proveen acceso en un dominio de aplicación dado a información que reside en fuentes heterogéneas y distribuidas, salvando al usuario de la complejidad de acceder y combinar esta información. Dada una serie de fuentes de información específicas, el *mediador* responde consultas sobre ellas, y el correspondiente plan de consulta involucra el acceso y combinación de información desde las fuentes necesarias. Un *Wrapper* es un componente que permite encapsular cada fuente, transformando los datos en un modelo común a todas las fuentes. De esta manera, el *mediador* proporciona un esquema global que acepta consultas del usuario y planea la ejecución de la consulta a través de las fuentes, de tal forma que si una consulta es ejecutada, ésta es pasada por el *wrapper* de una fuente específica, el cual traduce la solicitud en una forma entendible por la fuente, se la envía como una consulta y recibe los resultados. Finalmente, estos resultados son retornados al mediador y reunidos para devolver la respuesta a la consulta inicial del usuario. Al conjunto de Wrappers y Mediadores cooperando entre sí, se les conoce como un *Sistema de Información Basado en Mediadores* -SIBM- [Leser, 2000].

El artículo está organizado de la siguiente manera: La sección 2, presenta como marco contextual una *Introducción a la Planificación de Consultas con Control de Calidad en Sistemas Mediadores de Información*; la Sección 3, una descripción de los *Trabajos Relacionados* como una aproximación al Estado del Arte; la Sección 4, *Conclusiones y Trabajos Futuros*, como una aproximación a la propuesta en curso para enfrentar el problema.

2. LA PLANIFICACIÓN DE CONSULTAS CON CONTROL DE CALIDAD

Internet está basado en un paradigma de búsqueda que hace difícil recuperar e integrar datos desde múltiples sitios y en su mayoría provee aplicaciones como motores de búsqueda y meta-buscadores que permiten a los usuarios buscar la información que necesitan, por medio de una búsqueda basada generalmente en recuperación de información sintáctica de textos. Debido a esto se han desarrollado una serie de propuestas que tratan de integrar un conjunto de diferentes fuentes especiali-zadas, para lo cual extraen, filtran y representan eficientemente la información obtenida de la Web, pero la mayoría están enfocadas principalmente a la cantidad de información recuperada y a la calidad de la consulta, medida ésta con criterios de tiempo y costos de ejecución para encontrar planes óptimos, como lo muestran los trabajos de Chen *et al.* y Ambite *et al.*. En los Sistemas de Información Web, entre los que se encuentran los SIBM, el principal factor de eficiencia que tienen en cuenta las estrategias de planificación de consultas no es el tiempo de respuesta, como en los Sistemas de Información Tradicional, sino la calidad de la información –IQ- de los resultados de las consultas. Varias investigaciones se han acercado a este tema con la exploración de criterios para responder consultas de usuario en SIBM [Naumann, 2000].

La *Planificación de Consultas* es el problema de encontrar una secuencia de acciones para la ejecución de una consulta a través de fuentes de información autónomas, heterogéneas y distribuidas [Naumann, 2000]. El proceso de planificación especifica el flujo adecuado de los datos y el orden en el cual se deben desarrollar las diferentes operaciones y algoritmos específicos para cada uno, buscando en el espacio de los posibles planes y comparando el costo de cada uno de ellos.

La *calidad de la información* es un tema que tiene mucha consideración en SIBM y es tópico de investigaciones sobre captura y modelamiento de la información. Sin embargo, pocas han aplicado esa calidad al proceso de planificación de consultas sobre la Web [Naumann *et al*, 2001]. En los SIBM, la calidad de la información no solo se relaciona con la calidad del proceso de planificación, sino también con la calidad de la respuesta a la consulta. Para lograr mejorar la calidad del resultado de una consulta y el desempeño de los algoritmos de planificación, es necesario tener en cuenta criterios de IQ en la formalización del modelo de planificación a seguir en un dominio específico. Los sistemas de integración de información normalmente se han enfocado hacia medir criterios de calidad relacionados con minimización de costos y han recibido poca atención otros que se relacionan con la calidad de la información de las respuestas a las

consultas, esto es, aspectos como la relevancia de las respuestas de acuerdo a las necesidades del usuario.

3. TRABAJOS RELACIONADOS

A continuación se discuten diferentes propuestas relacionadas con la Calidad en la Planificación de Consultas en SIBM, que han sido reportadas en los últimos años y han sido agrupadas en tres categorías: Optimización de Consultas Tradicional, Calidad del Proceso de Planificación de Consultas y Calidad de la Información en Planificación de Consultas Web.

3.1 Optimización de Consultas Tradicional

En la literatura de Bases de Datos, la optimización de consultas ha sido ampliamente estudiada. Un optimizador de consultas intenta encontrar la forma algebraica más eficiente de una consulta y escoger métodos específicos para implementar cada operación de procesamiento de datos. La *Optimización de Consultas* está basada en transformaciones de árboles de consultas para generar planes de consultas más óptimos, por medio de varios teoremas que describen la optimalidad de algunos tipos de planes bajo diferentes modelos de costos: longitud de los planes y costo computacional [Chu y Hurley, 1982]. La investigación desarrollada en esta sub-área, se enfoca hacia la optimización de la consulta basada en criterios de eficiencia: minimización de tiempos y costos de ejecución.

El objetivo de búsqueda del usuario ha cambiado con el movimiento de la Optimización de Consultas en Sistemas de Información Tradicional a los Sistemas de Información Integrados, pues el principal criterio de optimización ya no es la minimización de tiempos y costos de ejecución. En este nuevo ambiente el usuario ya no demanda la respuesta *correcta*, sino su mayor satisfacción con respuestas aproximadas; no demanda la respuesta *absoluta*, sino la respuesta que sea más relevante; no demanda una respuesta *completa* con todos los atributos, sino que esté satisfecho con ciertos valores que busca. El objetivo de 'encontrar una respuesta completa tan rápido como sea posible' ha cambiado al problema dual de 'encontrar la mejor respuesta posible dentro de restricciones de tiempo y costo'. El surgimiento de los Sistemas de Información Web y su creciente desarrollo, ha ampliado los problemas relacionados con la 'baja' calidad de la información, pero al mismo tiempo ha logrado una audiencia con un nuevo perfil de requerimientos [Naumann, 2001].

3.2 Calidad del Proceso de Planificación en SIBM

Muchos de los trabajos actuales enfrentan la planificación de consultas en SIBM, teniendo en cuenta la calidad del plan en términos de la selección de las fuentes y la eficiencia en términos de costos de ejecución de las consultas.

El *Information Manifold* [Levy *et al.*, 1996] y el *TSIMMIS* [Hammer *et al.*, 1995] enfocan la Planificación hacia la optimización basada en costos, donde primero un conjunto de planes recuperables son encontrados y luego se optimiza cada uno independientemente. En *HERMES* [Adali *et al.*, 1996], el mediador usa un lenguaje lógico para integrar un conjunto de fuentes de información, por medio de un sistema que incluye un algoritmo que reescribe reglas y transforma los planes que evalúan una consulta de usuario a una más costo-efectiva logrando selecciones de las fuentes, reordenando sub-objetivos en reglas y usando técnicas de almacenamiento de ellas.

El *proyecto GARLIC* [Tork Roth *et al.*, 1996] y [Roth y Schwarz, 1997] considera la optimización de costos para mediadores y evaluación de sub-consultas de las fuentes de información. La optimización procede aquí en tres etapas: selección de las fuentes, programación dinámica de los planes y localización de operadores para escoger el mejor plan; de esta forma la optimización se basa en estos aspectos y no se consideran aspectos en tiempo de ejecución, ni optimización semántica que permita mejorar la calidad de la consulta y sus resultados.

El sistema *SAGE* [Knoblock, 1996], considera la calidad del plan, soportándola con la propuesta de intervención entre planificación y ejecución. El *sistema OCCAM* [Kwok y Weld, 1996] es un planificador para recuperación de información en dominios distribuidos y heterogéneos que se enfoca principalmente en el problema de la selección de

las fuentes relevantes para la consulta, más no del procesamiento de la consulta como tal, ni de la medición de la calidad de la información.

Un marco de trabajo más relacionado con el problema y que provee buenos resultados, se presenta en el estudio desarrollado por [Ambite, 1999] donde se propone el paradigma de la ***Planificación por Reescritura*** –PbR-, que combina la selección de las fuentes y la optimización de la consulta basada en costos. El paradigma de PbR es diseñado para tratar la eficiencia de la planificación y la calidad del plan, además de los beneficios de la independencia del dominio, lo que lo hace especialmente adecuado para el dominio de *planificación de consultas,* pues puede tratar cientos de fuentes y planes de consulta grandes y produce planes de buena calidad independiente del dominio. Aunque otras propuestas mejoran la eficiencia de la planificación con simples métricas de costo (como el número de pasos), ésta se caracteriza por mejorar la calidad del plan utilizando algoritmos de selección de fuentes que permiten asegurar la calidad de ellas, sumado a la optimización basada en costos.

En [Ives, 2002], se proponen y evalúan un conjunto de técnicas para procesamiento de consultas adaptativas, es decir, que se adaptan a su medio de ejecución y permiten al procesador de la consulta reaccionar a las condiciones cambiantes o al conocimiento que va creciendo en tiempo de ejecución. Estas técnicas están basadas en: algoritmos adaptativos para ejecución de consultas que proporcionan rapidez y más eficiencia en el procesamiento de las consultas, nuevos operadores del álgebra relacional para tratar datos del ambiente Web que hacen frente a las variaciones en las velocidades de transferencia de los datos, y re-optimización de consultas para mejorar los planes escogidos. Este trabajo propone entonces que un plan de consulta debe ser escogido adaptativamente, de tal forma que el procesador de consulta escoja un plan de consulta inicial que continuará refinándose cuando se monitoreen sus costos de ejecución y sus estadísticas, de manera que el sistema considerará factores que varían sobre el curso de la ejecución, así como factores que permanecen consistentes pero son previamente desconocidos.

3.3 Calidad de la Información en la Planificación de Consultas

Las siguientes propuestas, manejan de manera incipiente la calidad de la información en SIBM, utilizando sistemas Multiagente y esquemas de representación de fuentes de información.

La propuesta de [Camacho *et al.*, 2002] plantea la Arquitectura Multiagente Distribuida *MAPWEB,* que trabaja con consultas que acepta desde los usuarios y como resultado produce posibles esquemas de soluciones por medio de técnicas de resolución de problemas de planificación y aprendizaje. Mediante esta arquitectura se busca resolver problemas complejos que requieran integrar información heterogénea procedente de diferentes fuentes Web empleando para ello la cooperación de agentes de planificación y agentes Web. El problema de esta propuesta en cuanto a la planificación de consultas es que no tiene en cuenta algún tipo de criterio que permita responder a las consultas con información de calidad, de acuerdo a las necesidades iniciales del usuario. Relacionan la calidad con el nivel de razonamiento que alcanzan los Agentes de Información cooperando, pero por ser una propuesta de propósito general, no es muy concreto el nivel de razonamiento alcanzado y en qué aspecto específico lo logra.

En [Knoblock *et al.*, 1998] se propone el *Sistema ARIADNE*, que es una arquitectura que permite eficientemente integrar múltiples fuentes mediante un modelo de datos común. Esta propuesta se construye basada en técnicas de representación de conocimiento, aprendizaje de máquina y planificación automatizada, y hace más rápido y económico construir nuevos agentes de información que acceden a fuentes web existentes y hace más fácil mantenerlos e incorporar nuevas fuentes cuando estén disponibles. ARIADNE es un sistema que extrae e integra datos desde fuentes Web semi-estructuradas, y permite a los usuarios crear rápidamente 'agentes de información' para la Web. Está más enfocado hacia los esquemas de representación y herramientas de modelamiento de las fuentes, que hacia el procesamiento de la consulta como tal, en donde igualmente la calidad de los planes se mide con métricas basadas en costos, pero se logra un nivel de calidad de la información manejada por las fuentes, con el uso

de ontologías subyacentes a cada fuente que permite proporcionar información relevante a las consultas.

Los siguientes autores en sus respectivos trabajos tienen el mayor acercamiento encontrado en la literatura al problema del manejo de la calidad de la información en la planificación de consultas en SIBM, así:

En [Naumann, 2000], se investiga la exploración de criterios de *calidad de la información -IQ* -para responder consultas de usuarios y discute qué criterios IQ son necesarios, cómo se pueden adquirir y ser usados para mejorar la calidad de los resultados y el desempeño de los algoritmos de planificación. Este autor plantea el hecho de la importancia que tiene la calidad de la información en los sistemas distribuidos a gran escala, luego de algunos trabajos desarrollados en el área y enfatiza en la ausencia de investigaciones que apliquen razonamiento sobre la calidad de la información en el área de la planificación de consultas. Como prueba de su aproximación, desarrolla un sistema que encuentra eficientemente los resultados de consulta óptimos, es decir, con alta calidad basado en los criterios de calidad definidos, aplicado sobre un meta-motor de búsqueda que usa motores de búsqueda existentes como sus fuentes de información distribuidas. Como criterios de calidad para este dominio incluyen completitud y frecuencia de actualización. El trabajo en general cubre todos los aspectos de la integración: desde la definición de fuentes, y criterios de IQ, modelamiento de los resultados y métricas de los criterios para este dominio, métodos para evaluarlos y algoritmos que hacen uso de estos resultados para producir resultados de consultas de alta calidad. Esta investigación es un buen marco de trabajo guía para el análisis de los criterios de calidad que pueden ser implementados en ambientes web, pues el dominio de aplicación es abierto, heterogéneo y distribuido. Adicional a los criterios analizados, hay otros que pueden presentar buenos niveles de calidad en la planificación de consultas, como los tenidos en cuenta por *Chen et al.* en el siguiente trabajo.

[Chen *et al.*, 1998] presenta una investigación sobre la calidad del procesamiento de consultas en la WWW, debido a muchos factores tales como tiempo de respuesta impredecible, resultados irrelevantes y datos no actualizados. Propone un método para el procesamiento de consultas controlando la calidad en este ambiente Web. Introduce parámetros de calidad que los usuarios pueden especificar cuando se introducen las consultas, al igual que funciones que son usadas para evaluar la bondad de estos parámetros y algoritmos de programación, planificación y ejecución.

El trabajo presentado en [Marotta *et al*. 2002], está enfocado hacia la **evaluación de calidad**. Se intenta describir el problema de gestión de calidad en SIBM, proponiendo una solución para evaluación de calidad, experimentando con algunas propiedades y su clasificación, basado en la diferencia entre requerimientos de calidad del usuario y calidad ofrecida por las fuentes. Propone un mecanismo para deducir la calidad ofrecida por estos sistemas, la cual propaga los valores de calidad de las fuentes a las vistas del usuario y también hace conversiones entre diferentes clases de propiedades de calidad. Estos autores muestran que las propiedades de calidad no son necesariamente las mismas, pues la visión de los usuarios es diferente a la de los administradores del sistema, especificando la calidad requerida y la calidad real.

4. CONCLUSIONES Y TRABAJOS FUTUROS

En este artículo se mostró la necesidad de la Planificación de Consultas con Control de Calidad en el desarrollo de Sistemas de Integración de Información en la Web, los cuales normalmente se han enfocado hacia medir criterios de calidad relacionados con costos de planificación, costos de ejecución y completitud operacional de los planes de consulta, donde el tratamiento de otros criterios que se relacionan con la calidad de la información de las respuestas a las consultas, ha recibido poca atención, esto es, aspectos como la relevancia de las respuestas de acuerdo a las necesidades iniciales del usuario y la exactitud de éstas.

Se convierte en una necesidad el orientar la planificación de la consulta en SIBM hacia la calidad de la información devuelta como respuesta a una consulta inicial. Para avanzar en este campo,

como base preliminar para posibles investigaciones, es importante tener en cuenta trabajos como los desarrollados alrededor de Criterios de Calidad de la Información -IQ- en [Chen *et al.*, 1998] y [Naumann, 2000], y de Evaluación de IQ en [Marotta *et al*., 2002], que cubren todos los aspectos de la integración de información: definición de fuentes de información, definición de calidad de información y criterios de IQ, métricas de los criterios para dominios específicos, modelamiento de los resultados, métodos para evaluarlos y algoritmos que hacen uso de éstos para producir resultados de consultas de alta calidad.

Es importante proponer soluciones para evaluación de calidad, experimentando con nuevas propiedades o criterios de calidad, pero teniendo en cuenta la diferencia entre 'calidad requerida por los usuarios' y la 'calidad ofrecida por las fuentes'. Lograr una caracterización formal y precisa de criterios de calidad de la información es un problema difícil de tratar, y es trabajo de algunas investigaciones como [Marotta *et al*., 2002] pero aún permanece sin resolver. De la misma forma, hay problemas abiertos relacionados con métricas, representación y manejo de criterios de calidad de la información, que pueden ser objeto de trabajos posteriores.

Como una primera aproximación para enfrentar el problema de encontrar planes de ejecución de consultas a fuentes de información web, que permitan el control de la calidad, se pretende desarrollar un Modelo de Planificación de Consultas orientado hacia el control de la calidad de la información. Se espera tratar criterios de calidad de la información relacionados con las fuentes, como completitud y frecuencia de actualización, y relacionados con la intervención del usuario, como relevancia de las respuestas. Se pretende además que el modelo permita una continua intervención del usuario especificando los parámetros de calidad que desee en cada consulta.

Es necesario que el modelo de planificación se enmarque en el campo de acción de los SIBM y se centre específicamente en la planificación de la consulta, con el fin de comprender la estructura de los elementos básicos que se requiere tener en cuenta y proponer un modelo de planificación que posibilite el manejo de la calidad de la información. Se espera que este modelo de planificación de consultas, además de razonar acerca de la calidad de la información que procesa, propicie un espacio más de discusión acerca de la pertinencia de hablar de 'calidad de información' dentro del proceso de planificación de consultas.

RECONOCIMIENTOS

El trabajo descrito en este artículo ha sido apoyado por los siguientes proyectos de investigación:

Tesis de Maestría "*Modelo de Planificación de Consultas con Control de Calidad en Sistemas de Información Basados en Mediadores*", apoyado por la Escuela de Sistemas de la Universidad Nacional de Colombia, Sede Medellín.

Tesis de Doctorado "*Modelo Distribuido y Cooperativo Basado en Agentes Ontológicos y de Planificación, para la composición Automática de Servicios Web Semánticos*", auspiciada por Colciencias, ICFES, ICETEX, Universidad Nacional de Colombia, Sede Medellín y el Banco Mundial, enmarcado en el programa de apoyo a la comunidad científica nacional en programas de Doctorado 2004.

REFERENCIAS

Ambite, J.L. y Knoblock, C., (1997). Planning by rewriting: Eficiently generating high-quality plans. En *Proceedings of the 14 National Conference on IA*, Providence.

Ambite J.L., (1999) *Planning by Rewriting*. PhD Thesis, University of Southern California.

Camacho *et al.*, (2002) MAPWEB: Cooperation between Planning Agents and Web Agents. En *Information & Security.*.

Chen *et al.*, (1998). Query processing with quality control in the World Wide Web.

Chu, W. y Hurley, P., (1982). Optimal query processing for distributed database systems. *IEEE Transactions on Computers*,

Ives, Z., (2002). *Efficient Query Processing for Data Integration*. PhD Thesis, University Washington.

Kambhampati, S. y Knoblock, C., (2003). Information Integration on the Web. En: *Guest Editors' Introduction. IEEE Intelligent Systems*.

Knoblock *et al.*, (1998). Modeling web sources for information integration. En *Proceedings of the 15th National Conference on IA*, Madison, WI.

Knoblock *et al.*, (2000). The ARIADNE approach to Web-Based Information Integration. En *International Journal of Cooperative IS*.

Leser, U., (2000). Query Planning in Mediator Based Information Systems. PhD. Thesis Universitat Berlin.

Levy *et al.*, (1995). Answering queries using views. En *Proceedings of the 14th ACM Symposium on Principles of DBS*, San Jose, California.

Levy, A., (2000). Logic-Based Techniques in Data Integration. Department of Computer Science. University of Washington, Seattle.

Marotta *et al*. (2002). Quality Management in Multi-Source Information Systems. Facultad de Ingeniería. Universidad de la República. Montevideo, Uruguay.

Naumann, F.,(2000).Quality-driven Query Planning. Dissertation Outline Humboldt-Universit at zu Berlin.

Naumann *et al*, (2001). Quality-driven Integration of Heterogeneus Information Systems. Humboldt-Universitat zu Berlin.

Naumann, F., (2001). From Databases to Information Systems– Information Quality Makes the Difference. IBM AlmadenResearchC.