Un modelo de planificación de consultas basado en la calidad, en sistemas de información en la web

Bell Manrique Lozada[1]

Jaime Alberto Guzmán Luna[2]

Resumen

Este artículo presenta los resultados finales obtenidos en el desarrollo de un 'Modelo de Planificación de Consultas con manejo de la Calidad de la Información', cuyo objetivo es determinar lo que implica tener en cuenta la calidad de la información en el proceso de planificar una consulta en sistemas de información basados en mediadores, como un tipo de sistema de información web. Se presenta una exploración de criterios de calidad de la información que se pueden tratar en la planificación de una consulta, se propone un modelo de calidad y con base en éste un modelo de planificación. En el tema específico de manejo de 'calidad de la información' en la planificación de consultas hay poca información en la comunidad científica, por lo que se considera muy importante que el análisis y la forma de abordar el problema, el modelo propuesto y los resultados presentados aquí, contribuyan positivamente en este campo, lo que redundará en el fortalecimiento del conocimiento al respecto y en la ampliación de la aplicación de este tipo de sistemas de información.

Palabras clave: Calidad de la Información, integración de información, mediadores, planificación de consultas, web semántica.

A query planification, quality based on web information system

Abstract

This paper presents the final results obtains in the development framework of a Query Planning Model driven Information Quality in Information Integration Systems, whose objective is to determine the necessity characteristics for considering the information quality in the query planning process in Mediators-Based Information Systems. A exploration of information quality criteria which could deal in query planning are proposed; and a quality model and based in it a query planning model is proposed too. In specific topic 'driven information quality' in the query planning, there are few information in the scientific community, hence is important that the analysis and the way of focusing the problem, the proposed models and the partial results showed in this paper, positively to contribute and to make stronger this investigation area.

Key words: Information quality, information integration, mediators, query planning, semantic web.

1. Introducción

El acceso integrado a información que reposa sobre múltiples fuentes de información es un problema importante en muchos dominios actuales. Como una manera de enfrentar este problema nacen los sistemas de Integración de Información en la web, cuya tarea es responder consultas de usuarios extrayendo y combinando datos desde múltiples fuentes. En este contexto se han desarrollado propuestas que integran un conjunto de fuentes especializadas, la mayoría enfocadas principalmente a la cantidad de información recuperada y a la calidad de la consulta medida con criterios de tiempo y costos de ejecución para lograr planes óptimos [6], [1].

En los sistemas de información web, el principal factor de eficiencia en el procesamiento de consultas no es el tiempo de respuesta, como en Sistemas de Información Tradicional, sino la calidad de la información (IQ) [20]. La IQ es un tema importante, tópico de diversas investigaciones sobre captura y modelamiento de la información, sin embargo, pocas investigaciones tratan de aplicar esa IQ al proceso de planificación [20]. Los sistemas de integración de información normalmente se han enfocado hacia medir criterios de calidad relacionados con minimalidad, costos de ejecución y completitud operacional de los planes y ha recibido poca atención el tratamiento de otros relacionados con la IQ, como la relevancia de las respuestas.

Este trabajo presenta los resultados obtenidos en el marco del desarrollo de un Modelo de Planificación de Consultas con manejo de Calidad de la Información, que pretende determinar lo que implica tener en cuenta la IQ en el

[1] Docente Programa Ingeniería de Sistemas, Universidad de Medellín, miembro del grupo de investigación SintelWeb.

[2] Docente de la Escuela de Sistemas de la Facultad de Minas, Universidad Nacional de Colombia, miembro del grupo de investigación SintelWeb.

Artículo recibido en Abril de 2007, aprobado para publicación en Junio de 2007

proceso de planificar una consulta. En el tema específico de manejo de 'IQ en la planificación de consultas', hay poca información en la comunidad científica, por lo que se considera muy importante que el análisis y la forma de abordar el problema, el modelo propuesto y los resultados parciales presentados, contribuyan positivamente al fortalecimiento del conocimiento al respecto. Para tal efecto, este artículo está organizado de la siguiente manera: Sección 2, marco teórico y trabajos relacionados; sección 3, modelo de calidad propuesto; sección 4, proceso de planificación; sección 5, modelo de planificación propuesto; sección 6, validación y pruebas; y sección 7, conclusiones y trabajo futuro.

2. Marco teórico y trabajos relacionados

2.1. Sistemas de información basados en mediadores

Para manejar la integración de información se crearon componentes que interactúan entre sí y ofrecen un acceso integrado a los datos en un dominio específico, éstos son mediadores y wrappers. Los mediadores proveen acceso uniforme a información que reside en diferentes fuentes, protegiendo al usuario de la complejidad de acceder y combinar esta información. Un wrapper es un componente que permite encapsular cada fuente, actuando como un software que transforma los datos en un modelo común. Un conjunto de wrappers y mediadores cooperando, constituyen lo que se llama un Sistema de Información Basado en Mediadores –SIBM- [12]. En los SIBM, 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 heterogéneas [18]. En los SIBM, la IQ no solo se relaciona con la calidad del proceso de planificación, sino también con la calidad de la respuesta a la consulta y se puede medir con criterios como: solidez, completitud operacional y conceptual, relevancia y exactitud.

2.2. Calidad de la información

El objetivo de búsqueda del usuario ha cambiado con el movimiento de las bases de datos tradicionales a los SIBM, tal que el objetivo ya no es encontrar una respuesta completa tan rápido como sea posible, sino encontrar la mejor respuesta posible dentro de ciertas restricciones de tiempo y costo [19]. El tema de la IQ ha recibido mucha atención en los últimos años como resultado de la creciente importancia del valor de la información como ventaja competitiva. Investigaciones sobre IQ [2] proporcionan muchas perspectivas describiendo calidad en cuatro formas generales: como excelencia, valor, conforme a especificaciones, o encontrar o exceder las expectativas de un usuario. Las primeras dos perspectivas de calidad son problemáticas, pues definir la calidad como excelencia es subjetivo y no proporciona una guía para su mejoramiento o control, y como valor está sujeta a confusiones entre excelencia y costos. Las dos últimas vistas conforme a especificaciones y encontrar o exceder las expectativas del usuario, sí pueden ser definidas y medidas [9].

2.3. IQ en planificación de 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 [18], se investiga la exploración de criterios de IQ para responder consultas en SIBM y discute qué metadatos son necesarios, cómo pueden ser adquiridos y utilizados para mejorar la calidad de los resultados de la consulta y el desempeño de los algoritmos. [6] presenta una investigación sobre la calidad del procesamiento de consultas en la WWW, y propone un método para el procesamiento de consultas controlando 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 [17], está enfocado hacia la evaluación de calidad. Se describe el problema de gestión de calidad, se propone una solución para evaluación de calidad experimentando con algunas propiedades y su clasificación. Presentan un mecanismo para deducir la calidad ofrecida por el sistema, 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.

3. Modelo de calidad propuesto

En general no existe un consenso a la hora de definir y clasificar las características de calidad que debe presentar un producto, en este caso una fuente de información. Para lograr determinar las propiedades de calidad a tener en cuenta en el modelo, esta propuesta tiene en cuenta la terminología utilizada para productos software, en donde una característica de calidad de un producto es un conjunto de pro-

piedades mediante las cuales se evalúa y describe su calidad. En este trabajo, se llama criterio a una propiedad de calidad a la que puede asignársele una métrica, la cual es un procedimiento que examina un componente y produce un dato simple, un símbolo o un número.

Un modelo de calidad [4] es el conjunto de características y sub-características y de cómo éstas se relacionan entre sí. El objetivo principal es detectar los criterios que pueden describir la IQ en cada fuente, teniendo en cuenta la información que suministran y que facilitan su valoración por el planificador. La definición de los criterios fue estrictamente teórica, es decir, producto de una revisión de literatura sobre la IQ asociada a fuentes de información Web. Entre estos trabajos, se encuentran: [28], [3] en Georgia University, [26] en Dakota State University, [11] por MIT, UC, Berkerley y Boston University, MA y USA y clasificación de criterios por [21], [7].

3.1. Selección y definición de criterios

Como una primera aproximación a criterios de calidad que se pueden tener en cuenta en la planificación de una consulta [9], en el contexto de SIBM se pueden tratar criterios como: Completitud, cobertura, precisión, oportunidad, privilegios, disponibilidad, conectividad, objetividad, relevancia, reputación, seguridad, entendibilidad y valor agregado. Teniendo en cuenta la clasificación de criterios propuesta en [9], para nuestro propósito se determinaron tres niveles o clases de criterios de IQ, con el fin que cada uno de ellos se relacione con una fase del proceso de planificación. En esta clasificación, la IQ es influenciada por: La percepción del usuario, la información en sí misma y el proceso de acceder la información. De tal forma que se puede hablar de IQ en: El Usuario, pues es él quien decide si alguna información es cualitativamente buena o no; el Proceso de la consulta, que contiene criterios relacionados con el proceso de acceder la información; y la fuente, pues para muchos criterios la fuente en sí misma es el origen de los criterios de IQ. En cada una de las fases de planificación relacionadas con usuario, proceso y fuente, se definió un modelo de calidad compuesto por un criterio, una descripción de criterio y una métrica asociada. Los criterios definidos, son: Relevancia, cobertura, precisión, confiabilidad (reputación) y tiempo de respuesta.

Para obtener una definición aceptable de calidad, normalmente se utilizan los conceptos de métrica y medida [8]. Según [5] una métrica es una medida cuantitativa del grado en que un sistema, componente o proceso posee un atributo dado. Para este contexto de trabajo, una métrica para un criterio es la medida cuantitativa del grado en que una fuente posee cierto criterio, tal que examina una fuente y produce un dato (símbolo: -SI, NO –F, V o número). Siguiendo la estructura de una métrica [19], ésta debe contener: Definición, unidad, escala y fuente. Según el número de fuentes que maneje el SIBM y el número de criterios IQ definidos, se tiene una matriz que contiene los vectores asociados a cada fuente [20], conteniendo pesos asignados según el número de criterios.

Teniendo en cuenta el número de fuentes que maneje el SIBM y el número de criterios IQ definidos, se manejará una matriz como la mostrada en la Tabla I, en donde las columnas representan el número de fuentes asociadas al sistema y las filas el número de criterios IQ. Los valores en cada casilla representan el valor asociado a cada criterio para cada fuente, según métrica determinada. … …

Tabla I. Matriz Valores IQ del SIBM

Fuentes	Fuente *1*	Fuente *2*	Fuente *n*
Criterio *1*	0.5	1	...
Criterio *2*	0	0.9	...
Criterio *n*	...	...	...

Como producto de esta matriz, se genera un vector asociado para cada fuente de información, con pesos asignados a él teniendo en cuenta el número de criterios definidos. Por ejemplo si una fuente A tiene tres criterios asociados, su vector asociado será $Fuente(Aiq) = \{5,0,3\}$, en donde 5, 0 y 3, son los pesos asignados a cada criterio

Tabla II. Funciones de mezcla de Criterios IQ

Criterio	Función de Mezcla	Breve Explicación
Disponibilidad	*A*B*	Probabilidad que los dos sitios estén disponibles
Precio	*A+B*	Las dos consultas deben ser pagadas
Tiempo Respuesta	*Max[A,B]*	Las dos son procesadas en paralelo
Exactitud	*A*B*	Probabilidad que los dos sitios no contengan un error
Relevancia	*A*B*	Probabilidad para coincidir la unión
Completitud	*A+B–A*B*	Probabilidad que el sitio A o B no tenga valores nulos
Cantidad	*A+B*	Todos los atributos que deben ser tratados

3.1.1. Funciones de mezcla de pesos IQ.

Para generar un valor IQ general en el proceso de integración de fuentes, es decir, al concatenar dos o más valores de dos o más fuentes, es necesario utilizar funciones de mezcla como las que se muestran en la Tabla II, propuestas en [19].

Para nuestro caso, se asume que en el modelo de calidad propuesto, las fuentes de información del sistema son independientes, tal que se encuentra que siempre hay un atributo común en cada una de ellas. Para cada criterio del modelo de calidad se tiene que:

(i) Cobertura y Precisión**:** Se determina la cobertura y precisión general sumando los valores IQ y restando la posible coincidencia entre la nueva fuente y la existente. Debido a la suposición de independencia de las fuentes, se asume que se puede cuantificar esta coincidencia como el producto (multiplicación) de los dos valores.

$$Cobertura(cob)\ y\ Precisión(pre): [A + B - A * B] \quad (1)$$

(ii) Relevancia y Reputación: Se determina la relevancia y reputación general multiplicando los valores IQ de cada fuente. Debido a la suposición de independencia de las fuentes, se asume que se puede cuantificar esta coincidencia como el producto de los dos valores.

$$Relevancia\ (rel)\ y\ Reputación\ (rep): [A * B] \quad (2)$$

(iii) Tiempo de Respuesta: Se determina el tiempo de respuesta general, seleccionando el valor máximo IQ de tiempo de respuesta entre las dos fuentes a ser mezcladas.

$$Tiempo\ de\ Respuesta\ (tir): Max\ [A,B] \quad (3)$$

4. Proceso de planificación

Con el fin de determinar cómo debe ocurrir el proceso de planificación de la consulta, primero se realizó una exploración de algoritmos de planificación de consultas para determinar cuál de ellos podría ser el más indicado para ejecutar el modelo de calidad propuesto. Luego se determinó el algoritmo genérico, con los pasos necesarios para el manejo de la IQ.

4.1. Exploración de algoritmos

Han sido estudiados los siguientes algoritmos que permiten responder consultas usando vistas y que han sido previamente evaluados en [14], [15]. El algoritmo bucket, utilizado en el contexto del Sistema de Información Manifold [16]; el algoritmo de regla inversa, implementado en el sistema InfoMaster [25]; el algoritmo MinCon [24] y el algoritmo HiQA, utilizado en el proyecto HiQIQ [13].

En general, HiQA logra un mayor nivel de eficiencia debido principalmente a su capacidad de poda, pues las primeras ramificaciones pueden ser podadas y los planes óptimos alcanzados más rápido; se basa en el algoritmo Bucket, pero extiende su aplicación teniendo en cuenta valores de IQ por medio de los cuales poda la ramificación en el espacio de búsqueda. La diferencia entre el Bucket y el de regla inversa es que el primero computa las vistas relevantes tomando en consideración el contexto en el cual el componente aparece en la consulta, mientras que el Inverso no lo hace. La consulta re-escrita obtenida por el Inverso puede resultar en vistas que no son relevantes, sin embargo puede ser computado una vez y ser aplicado a cualquier consulta. La fortaleza del Bucket es que explota los predicados en las consultas para reducir significativamente el número de candidatos conjuntivos que necesitan ser considerados. El Inverso tiene la ventaja de ser modular, puede ser computado antes de tiempo y es independiente de una consulta específica.

Con el fin de determinar cuál de los algoritmos explorados puede aproximarse a las necesidades requeridas por el sistema para manejar la calidad de la información, se realizó un chequeo de las características que cumple cada uno de ellos, enmarcada dentro de los siguientes aspectos:

- Desarrollados para mediadores: Son algoritmos cuya primera versión ha sido desarrollada para aplicarse a SIBM.
- Consultas conjuntivas: las que utilizan una serie de operadores select-project-join.
- Características desempeño: resultados de una evaluación experimental de los algoritmos presentada en [24], basándose en tiempos de procesamiento/respuesta de cada algoritmo. Las unidades representan la posición que ocupan según el desempeño mostrado en la experimentación (1 menor, 5 mayor desempeño)
- Consultas vistas virtuales: Algoritmos que se pueden aplicar a sistemas que respondan a consultas sobre vistas virtuales y no solo sobre vistas materializadas como en el caso de los datawarehouse.
- Bondades IQ: Son algoritmos que tienen en cuenta algún tipo de criterio de calidad

de la información dentro de su procesamiento, o que ha tenido alguna adaptación dentro de una aplicación desarrollada.

Como resultados, teniendo en cuenta que el principal objetivo del algoritmo escogido debe ser el tratamiento de la IQ que maneja cada fuente asociada al sistema, se seleccionó el HiQA para comparar el tipo de criterios que ha tenido en cuenta en sus aplicaciones, que son: Cobertura, Precisión (trabajados dentro de la categoría Completitud) y Tiempo de Respuesta. Los criterios Relevancia y Reputación, definidos dentro del modelo de calidad propuesto no han sido trabajados en las aplicaciones de este algoritmo.

4.2. Algoritmo requerido

La tarea del planificador es crear un plan de ejecución de una consulta y enrutarlo a las fuentes respectivas, por medio de la división en subconsultas. El algoritmo adaptado para los propósitos de esta tesis se llamará algoritmo B&H, en razón a que dos de las estrategias utilizadas en él han sido tomadas de los algoritmos Bucket y HiQa, respectivamente, nombre que se utilizará desde este momento.

Un usuario envía una consulta en términos del esquema mediado, el cual es un conjunto de relaciones diseñadas para capturar los aspectos relevantes de la aplicación. Los datos, si embargo, están almacenados en las fuentes. Para lograr traducir las consultas de usuarios en consultas sobre las fuentes de datos, el sistema necesita una descripción de los contenidos de las fuentes. Una de las propuestas para especificar tales descripciones es describir una fuente como una vista sobre el esquema mediado, especificando cuáles tuplas pueden ser encontradas en la fuente. Dada una consulta U, el sistema primero necesita reformular U para referirse a las fuentes de datos/vistas.

4.2.1. Representación del algoritmo B&H

La representación de cada componente en el B&H, se describe a continuación como elementos de entrada al algoritmo de planificación. B&H, procede como indica la Figura 1.

Input:
Q= consulta = {x1, x2, ..., xn}
//Ejemplo: {Título(x), autor(y), Año(z)}
g =sub-objetivos, donde *g* E *Q*
//g1 = Título(x); g2=autor(y)
V = vistas = *{v1, v2, ... , vn}*,
para cada una de las fuentes del sistema
IQ_Vi = vector de IQ para cada vista contenida en *V*
IQ_q = vector de IQ especificado por el usuario en la consulta *u*
Output: plan consulta para *u*

```
1.  Para g=1 hasta n, hacer
2.      Para cada vista v
3.          encontrar todas las posibles combinaciones de vistas que
            cumplan con los atributos (sub-objetivos) solicitados en la
            consulta
4.      fin
5.      guardar cada combinación en vector P
    //guarda cada combinación de V para g
6.  fin
7.  iq_q = {a,b,c,d,e}
    //vector pesos criterios IQ consulta
8.  iq_vi = {a,b,c,d,e}
    //vector pesos criterios IQ vistas
9.  para cada P hallada
    //cada combinación de vistas
10.     hallar iq_p combinando sus vectores particulares (iq_vi),
                según las funciones especificadas en el modelo de calidad
11.     guardar cada vector iq_p generado para cada plan
12. fin
        //iq_p = vector iq para cada plan
13. para cada p del conjunto P de planes generados
14.     para cada iq_p
15.         comparar los vectores iq_q con iq_p
16.         escoger los pesos iguales o mayores que los de iq_q de
            acuerdo al rango de valores especificados en el modelo de
            calidad
17.     guardar en P
18. retornar P
```

Figura 1. Algoritmo Adaptado -B&H-.

4.2.2. Comportamiento del algoritmo B&H

Se recibe una consulta y un vector con 4 pesos asignados por el usuario como criterios de búsqueda adicionales (de acuerdo a los 4 criterios de calidad definidos en el prototipo del sistema). La consulta tiene entonces unos metadatos asociados, que vienen a ser los sub-objetivos. Cada fuente tiene una vista asociada, es decir la especificación de los metadatos de los cuales posee información, en el mismo formato de la consulta. Según el metadato de cada sub-objetivo, se escogen las vistas que contengan los metadatos asociados. Para cada subobjetivo se buscan las vistas asociadas, y se van guardando en Pi. Cada vista también tiene un vector IQ asociado que contiene los pesos para los criterios de calidad definidos para ella. De la misma forma, la consulta tiene un vector IQ definido por el usuario. Teniendo en cuenta las funciones de mezcla de criterios de IQ definidas en el modelo de calidad del prototipo del sistema, se generan los vectores IQ para cada uno de los planes generados en el paso anterior (posibles com-

binaciones de vistas), logrando los vectores IQ asociados a cada plan. Para cada plan contenido en Pi, se compara el vector IQ asociado, con el asociado a la consulta y se escoge el que sea igual o superior al vector-iq-consulta, igualmente con base en las especificaciones de rangos y valores del modelo de calidad. Se guarda el plan escogido en P, para luego retornarlo.

Para la estructuración del modelo de planificación, se tomó como base la Arquitectura de Planificación propuesta en [10]. Como se muestra en la Figura 2, los componentes principales que contiene el proceso de planificación son: un Planificador, donde se realiza la fusión de las fuentes; un Componente Soporte de Fuentes, en donde se almacena información relacionada con características de cada fuente del sistema; un Monitor, de la planificación y ejecución; y un Ejecutor de las Consultas, quien finalmente devuelve una respuesta a la consulta inicial del usuario.

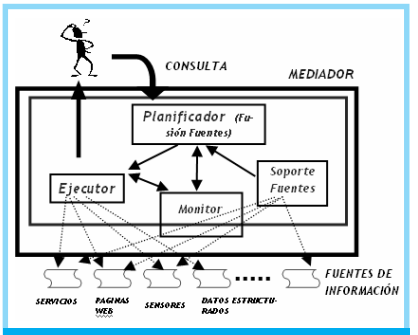

Figura 2. Arquitectura base SIBM.

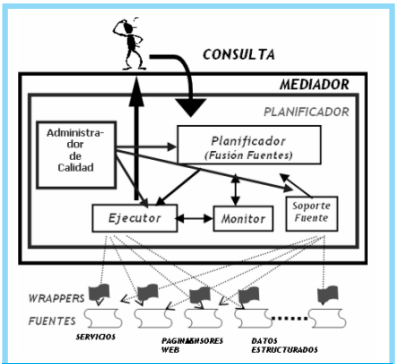

Figura 3. Arquitectura general propuesta

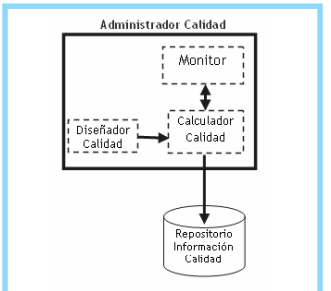

Figura 4. Arquitectura interna del módulo administrador de calidad.

Teniendo en cuenta el modelo de calidad propuesto y los pasos generales que debe seguir el algoritmo de planificación, la Figura 3 muestra una primera aproximación a la arquitectura del modelo de planificación con control de calidad de la información en SIBM. Se propone que el Planificador tenga un componente adicional 'Administrador de Calidad' que le proporcione información acerca de los criterios de IQ definidos por el administrador o diseñador del sistema y que al mismo tiempo permita realizar acciones de edición, monitoreo, registro y almacenamiento de este tipo de información. Los componentes internos de este módulo administrador de calidad, se detallan en la Figura 4.

El Módulo Administrador de Calidad es el componente es el que ha sido adicionado al planificador. Éste es realmente una parte del planificador, que permite conocer cómo calcular la calidad de las consultas a ciertos recursos que maneja. El planificador consulta las ontologías para recuperar los nombres de las fuentes que satisfacen una consulta, genera posibles subconsultas y solicitudes a cada una de las fuentes para evaluar el costo de consultar las respectivas subconsultas y cada una de las fuentes invoca sus módulos de calidad y estima el peso de su calidad total de acuerdo a los criterios de la consulta.

Desde la Interfaz de Generación de la Consulta, el proceso se lleva a cabo de la siguiente manera: Por medio de una interfaz el usuario escribe las palabras clave de búsqueda al igual que los pesos para los diferentes criterios IQ; con ayuda de estos valores, el mediador luego decide cuáles y cuántas fuentes usar; por los valores en el vector de pesos el usuario especifica la importancia individual de los diferentes criterios IQ; con base en este vector de entrada, el planificador compara con los valores IQ de cada fuente, guardados en el Repositorio de Calidad, para tomar la decisión de cuáles fuentes consultar. Los componentes, son:

(i) Diseñador de Calidad. Permite que el administrador o diseñador de calidad del sistema, determine cuáles criterios de calidad de la información van a ser tenidos en cuenta en el SIBM.

(ii) Calculador de Calidad. Determina el valor asociado a los criterios de cada fuente, teniendo en cuenta la información que recibe del Diseñador de Calidad (criterio, métrica, escala, etc.).

(iii) Repositorio de Información de Calidad. Componente de almacenamiento de los cál-

culos y valores de calidad asociados a cada fuente del sistema, que recibe del Calculador de calidad.

(iv) Monitor. Controla y monitorea que la planificación de la consulta se lleve a cabo con base a la información del módulo Administrador de Calidad.

El Nivel de mediación es el representado en general por la Figura 3 y es el núcleo del sistema, en el cual se enfoca la tarea de planificación. Consita de tres capas: La capa de presentación, encargada de encaminar, transformar y validar las consultas y los resultados obtenidos entre las aplicaciones externas y la capa de integración. La capa de datos, resuelve las heterogeneidades de las fuentes de información y encamina las consultas y resultados entre la capa de integración y las fuentes de información. El modelo de planificación propuesto se enmarca dentro de la Capa de Integración. La Capa de Integración, es donde se procesan y distribuyen las consultas procedentes de la capa de presentación y se integran los resultados que devuelve la capa de datos y se compone de:

(i) Repositorio de Información Local. Contiene información integrada obtenida a partir de las consultas realizadas por las aplicaciones. Permitirá resolver consultas sin acudir a las fuentes de información.

(ii) Sistema de procesamiento de consultas. Realiza las siguientes funciones:

- Recibe las consultas enviadas por el transformador de consultas
- Optimiza las consultas recibidas en función de la información almacenada en el catálogo (vectores de calidad)
- Realiza el particionamiento de la consulta global en un conjunto de consultas sobre cada fuente donde se guarda información relativa al objeto consultado.
- Encamina las sub-consultas hacia los Wrappers que realizan las funciones de intermediarios con las fuentes.
- Integra la información almacenada en el repositorio local con la información recibida.
- Envía la información integrada de salida al gestor de seguridad y encaminamiento.

(iv) El catalogo, que almacena:

- El esquema global formado a partir de los esquemas locales de las fuentes
- Metadatos, que representan relaciones entre los distintos objetos del esquema global, jerarquías de objetos, reglas de validación, relaciones entre los esquemas componentes, etc.
- Información relativa a los usuarios, permisos, privilegios y grupos definidos para el sistema de integración.

5. VALIDACIÓN DEL MODELO

5.1. Definición del prototipo

El dominio de aplicación del prototipo es un SIBM con documentación sobre libros, como una primera aproximación para el desarrollo de una biblioteca digital. A continuación se ilustran los principales componentes de la arquitectura.

Fuentes de información: En el prototipo, el SIBM tiene la siguiente relación global:

libros (ISBN, autor, título, publicista, año, país)

y comprime las siguientes vistas, relacionadas respectivamente con las tres fuentes de información:

El sistema integra información de tres fuentes de libros, como muestra la Tabla III, las fuentes 1 y 2 tienen información acerca de títulos y sus autores; y la fuente 3 tiene información acerca de títulos y el país y año de su publicación. Suponga un usuario que quiere encontrar los títulos de los libros escritos en el 2000 y publicados en la librería nacional. La respuesta puede ser obtenida tomando la unión de los siguientes joins: v1⋈ v2 v3 y v2 v3 y desarrollando una selección Año = 2000 y luego una proyección sobre el atributo publicista = nacional.

Tabla III. Vistas del Sistema de Prueba

Fuente	Vistas (Contenidos)
S_1	V_1 (ISBN, título, autor)
S_2	V_2 (isbn, título, autor, publicista)
S_3	V_3 (ISBN, título, año, país)

Las consultas al sistema son formuladas en términos de las relaciones del modelo-global, liberando al usuario de tener que conocer el vocabulario específico usado en cada fuente. Por ejemplo, imagine un usuario interesado en encontrar libros japoneses y sus reseñas, junto con sus precios asociados. Normalmente, el usuario debería tener acceso a cada una de las fuentes de libros japoneses individualmente, recuperar todos los que puedan ser relevantes y luego buscar la reseña para ellos. En un sistema como éste, el usuario puede enviar una consulta como:

q (titulo, autor, reseña, pais, precio): book (titulo, año, precio, autor) y reseña (titulo, año, reseña) y BookFg(país) y País = Japón

Dada una consulta específica, el sistema usa las descripciones de las fuentes para generar un plan para responderla. La idea del sistema es generar un plan que pode las fuentes de información de acuerdo a las especificaciones de calidad declaradas por el usuario. Para responder esta consulta, el algoritmo primero decide cuáles fuentes son relevantes, luego encuentra todos los posibles planes considerando las combinaciones relevantes y por último poda las posibilidades y escoge la combinación de fuentes que mejor se adapte al vector de IQ declarado por el usuario.

Agregando pesos de usuario: Normalmente no todos los usuarios esperan resultados de una consulta con igual importancia de calidad. Es posible encontrar consultas en las que el usuario esté interesado en ciertos atributos con un alto grado de precisión, y menor importancia en la actualidad o frescura de los datos; o con atributos actualizados hace poco tiempo, pero que no tengan una representación concisa; o atributos que sean relevantes a las necesidades de búsqueda, sin importar el tiempo de respuesta. Como una manera de enfrentar este problema, se propone que el usuario defina unos pesos (valores) para cada criterio de calidad, dependiendo de sus necesidades específicas de búsqueda. De esta forma un usuario puede preguntar al sistema por libros de cierto autor con su respectiva crítica, estando interesado en la información más actual (mayor frecuencia de actualización), sin importarle el tiempo de respuesta, tal que define unos valores: Actualidad = 10 y Tiempo de Respuesta = 4, y así mismo para el resto de criterios. El usuario puede dar un peso w_i para cada criterio IQ definido en el sistema. El peso debe ser dado dentro del rango común definido en el Modelo de Calidad. En la mayoría de los casos se utiliza el rango intuitivo de 0 (menor importancia) a 1 (mayor importancia).

5.2. Ejecución fases planificación

Crear buenos planes de ejecución para una consulta de usuario en un SIBM involucra un espacio de búsqueda de todos los planes y un modelo de valoración para comparar los planes con otros. En este trabajo se define el espacio de búsqueda en nuestro ambiente de SIBM como el conjunto de todos los planes que responden la consulta de usuario en una manera semánticamente correcta. Sin embargo, estos planes producen diferentes resultados pues pueden involucrar diferentes fuentes. Así, en general, más de un plan debería ser ejecutado para obtener una respuesta que sea tan completa como sea posible. Debido a la heterogeneidad de las fuentes de información en calidad y en costos, nos enfocaremos únicamente en el modelo de calidad para evaluar los planes.

Se propone que el proceso de planificación se desarrolle en fases, con base en las fases del algoritmo de planificación adaptado, manejando la IQ:

(i) Fase 1: Selección de la Fuente. En la primera fase se usan las vistas del sistema y los subobjetivos de la consulta para generar todos los posibles planes, es decir todas las combinaciones de las fuentes que proporcionen la información buscada, y así se establece el espacio de búsqueda para la siguiente fase.

(ii) Fase 2: Creación de Planes. En la segunda fase se explora el espacio de búsqueda completo usando los criterios definidos y los vectores asociados a cada fuente respecto a ellos y se escoge el mejor plan de ejecución. La entrada adicional para esta fase es el conjunto de vectores de IQ para los criterios descritos en el modelo de calidad. Algunos de estos pesos son predeterminados, mientras que otros son funciones de la consulta y las preferencias de usuario.

(iii) Fase 3: Definición Plan para Ejecución. En la tercera fase se escoge el plan que más se acerque a los requerimientos IQ de búsqueda del usuario, el cual es pasado para su ejecución.

5.3. Implementación del prototipo

El sistema prototipo diseñado está implementado en Java como una aplicación Web, usando una arquitectura cliente-servidor. El corazón del sistema es la aplicación en Java comunicándose con una base de datos relacional MySQL vía JDBC, con la cual se ha probado el prototipo. El esquema de la base de datos es sencillo, conteniendo la información relacionada a cada vista, es decir, los valores para cada criterio de IQ y los términos del modelo relacional global de los cuales posee información.

Como núcleo del prototipo es usado el algoritmo adaptado y presentado anteriormente, para encontrar las vistas asociadas a cada uno de los términos dentro de los resultados intermedios obtenidos desde la consulta de usuario.

5.4. Evaluación empírica-comparación

Para ilustrar la escalabilidad y desempeño del prototipo se desarrollaron pruebas del sistema. Las entradas consisten de consultas artificiales enviadas al sistema por un usuario normal de prueba, diseñadas para probar los efectos sobre el desempeño de responder las consultas teniendo como base diferentes preferencias de calidad del usuario. Luego, se desarrollaron tres casos de prueba del prototipo del sistema, que permite ver el comportamiento del modelo de calidad, del modelo de planificación y del algoritmo adaptado. Retomando, las vistas asociadas a cada fuente de información son: V1 (isbn, título, autor), V2 (isbn, título, autor, publicista) y V3 (isbn, título, año, país).

Para ilustración, en el CASO_1 un usuario envía una consulta Q preguntando por todos los títulos de los libros de los que el sistema posea información, junto con su autor, del año 2002. Además especifica sus intereses en cada criterio de IQ, así: Cobertura(0.9), Precisión(0.9), Reputación(0) y Tiempo de Respuesta(4.0). De esta forma, los pesos del usuario reflejan la importancia que le da el usuario a cada criterio o propiedad de calidad. En este caso, el usuario expresa por ejemplo más interés en la cobertura y precisión de la fuente que en su reputación: Q = [título, autor, año(2002)]; iq_q = {0.9, 0.9, 0, 40}.

Primero, para cada uno de los sub-objetivos de la consulta -título, autor y año-, se determina el conjunto de vistas que contiene cada sub-objetivo de la consulta, y se va almacenando este conjunto en un bucket o cubeta. Para los tres sub-objetivos de la consulta, se construyen entonces las siguientes cubetas: C1 (título)= {V1, V2, V3}, C2 (autor) = {V1, V2} y C3 (año) = {V3}.

Segundo, se enumera el producto cartesiano de todas las cubetas y se chequea si se satisface la consulta y si puede ser minimizada (vistas que son redundantes). Después de esto, los siguientes planes se generan, que producen potencialmente el conjunto de tuplas correctas para la consulta Q.

$P1 = V1 \bowtie V3 \quad P2 = V1 \bowtie V2 \bowtie V3$

$P3 = V2 \bowtie V1 \bowtie V3 \quad P4 = V2 \bowtie V3$

$P5 = V3 \bowtie V1 \quad P6 = V3 \bowtie V2$

El algoritmo también hace un filtro de aquellos planes en los cuales intervienen las mismas vistas, pero en diferente orden, para evitar cómputos dobles en el siguiente paso.

Cuando se han hallado todos los planes potenciales, se definen los vectores IQ asociados a cada plan, con base en las funciones de mezcla definidas en el Modelo de Calidad. La selección del plan procede entonces en tres pasos: 1) se determinan los vectores IQ para cada vista; 2) los vectores IQ de las vistas son mezclados para obtener una calidad general; y 3) el vector general es usado para comparar con el vector IQ de la consulta. Los pesos de cada criterio IQ para cada una de las vistas, se muestran en la Tabla IV.

Tabla IV. Pesos Criterios IQ, Case 1

	Vista1	Vista2	Vista3
Cobertura	0,9	0,7	0,5
Precisión	0,5	0,7	0,9
Reputación	0	0,8	1
Tiempo	30	45	80

Siguiendo la idea de modelos de costos en DBMSs, para el prototipo del sistema se ha diseñado un modelo de calidad para calcular el vector de pesos general o total de un plan. Dado que solamente se consideran operadores join, un plan puede representarse como un árbol binario con las vistas como hojas y los operadores join como nodos internos, como se muestra en la Figura 5. El vector IQ para un nodo interno es calculado como una combinación de los vectores IQ de sus nodos hijos derecho e izquierdo. La Figura 5 muestra el árbol de planificación para P2. En cada criterio, los vectores IQ son computados con el operador de mezcla, con base en las operaciones de mezcla definidas en el modelo de calidad. Teniendo en cuenta que las funciones de mezcla son conmutativas y asociativas, un cambio en el orden de ejecución del join dentro del plan no tiene efecto sobre el vector IQ resultante.

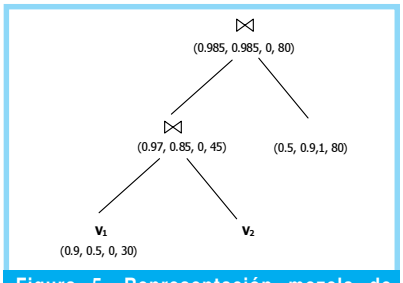

Figura 5. Representación mezcla de vectores IQ en nodos join del plan P2

Los tres planes, siguiendo este mismo procedimiento, tendrán los pesos de la Tabla V, asociados a cada criterio de IQ:

Tabla V. Vectores Planes Caso_1

	P1(V1,V3)	*P2(V1,V2,V3)*	*P3(V2,V3)*
Cobertura	0,95	0,985	0,85
Precisión	0,95	0,985	0,97
Reputación	0	0	0,8
Tiempo	80	80	80

Finalmente se comparan los vectores IQ de cada plan con el vector IQ de la consulta definido por el usuario, para escoger el que más se aproxime a éste. Cuando los criterios son positivos, como precisión, cobertura y reputación, se escoge el valor igual o mayor al solicitado en la consulta, o en su defecto el más cercano. Cuando los criterios son negativos como tiempo de respuesta, se escoge el menor valor o el que se acerque más por abajo al de la consulta, pues en estos casos el peso más alto corresponde al de peor calidad.

De esta manera, inicialmente se desecharía P3, pues: su peso de cobertura es menor a los demás y al de la consulta, su peso de reputación es alto y para el usuario es irrelevante, y el tiempo de respuesta es el doble del anotado en la consulta. Quedan los planes P1 y P2, en donde se escoge el P2 por tener dos de los pesos de importancia más altos que los solicitados y corresponden a criterios positivos.

6. Conclusiones y trabajos futuros

La planificación de consultas para integración de información ha recibido considerable atención en la comunidad de Bases de Datos y Sistemas de Información Tradicional, durante los últimos años. La planificación de consultas depende ampliamente de las formas como las fuentes son modeladas con respecto al esquema global del dominio de aplicación, tal que la calidad de la información procesada en la planificación y devuelta en un plan potencial, depende directamente del modelamiento conceptual realizado con cada fuente que proporcione información al sistema y con cada relación del esquema global.

Cuando se habla de planificar consultas de usuario contra múltiples fuentes de información integrada, heterogénea y distribuida, se habla de una nueva visión de la calidad de la información. Además de las dificultades tradicionales como conflictos entre esquemas de datos e interfaces heterogéneas, el almacenamiento de datos comunes en diferentes fuentes, también el usuario tiene demandas diferentes hacia el ambiente Web, enfocado al tratamiento de la calidad de la información. Es en este punto hacia donde se ha direccionado este trabajo, presentando un modelo para un SIBM basado en una relación global, en criterios de IQ y en un algoritmo de planificación que toma las decisiones soportado en ellos, para determinar cuáles fuentes son más apropiadas para responder una consulta de usuario.

Es evidente la necesidad de la Planificación de Consultas con Control de Calidad en el desarrollo de Sistemas de Integración de Información en la web, tal que es urgente orientar la planificación de la consulta en SIBM hacia la IQ devuelta como respuesta a una consulta inicial.

La visión de la web semántica está en auge en estos momentos permitiendo interoperabilidad entre agentes y apoyando a los usuarios humanos a localizar información relevante a sus necesidades de búsqueda y darle sentido a la información. Este contexto conlleva a que información semántica pueda ser usada en diferentes formas para mejorar los procesos de responder consultas y como consecuencia importante, la posibilidad que la Web sea consultada eficiente y directamente. El desarrollar trabajos sobre la Web semántica proporciona un nuevo y potencial contexto en el cual pueden ser aplicados los resultados de investigaciones como ésta.

El prototipo del sistema toma entradas de un usuario en forma de una consulta y un vector de IQ requerido y retorna una combinación de vistas desde un conjunto de Fuentes de información, el cual instancia la consulta de entrada con información específica del dominio. El sistema presenta una solución diferente en cuanto a la estrategia de planificación, combinando varias sub-estrategias en una. Utiliza una estrategia de similaridad de sub-objetivos (términos de la consulta) con vistas asociadas al sistema, una estrategia de generación de posibles combinaciones de las vistas de acuerdo a una consulta de usuario (planes), una estrategia de combinación de vectores IQ específicos de cada vista, en un vector total, y por último una estrategia de comparación de valores IQ de usuario y de las vistas.

Como trabajo futuro, se proyecta desarrollar SIBM bajo el modelo propuesto, tal que permita que un usuario a través de una interfaz de consulta, pueda buscar la información sobre la cual está interesado en un dominio específico, con particularidades específicas de calidad. La idea del sistema futuro es que sea interactivo y que los valores IQ asignados por el usuario ayuden a reformular y entender la consulta en términos de los conceptos manejados.

Como un aspecto importante para mejorar en el modelo propuesto, está la evaluación y/o retroalimentación del usuario de los resultados alcanzados con los pesos de IQ asignados por la fuente de información o ingresados al sistema por entradas de un experto o técnicas

de manipulación de los datos. El objetivo ideal es que la intervención del usuario en la evaluación de la información devuelta como respuesta a consultas iniciales, se vaya almacenando progresivamente, tal que sirva para la toma de decisiones posteriores por parte del planificador. De esta manera sería evidente el 'razonamiento' llevado a cabo por el planificador.

Se planea también trabajar en la optimización de los planes, sin dejar a un lado la calidad de la información que se maneja. Se pretende desarrollarlo como un paso de post-optimización, para encontrar el mejor orden de los operadores Join de los planes escogidos, sin influenciar negativamente sus vectores de IQ.

Trabajo futuro también incluirá cooperación de las fases de creación de los planes y selección de los planes, para mejorar el tiempo de planificación, tal que los resultados encontrados sean de buena calidad, al igual que los planes, logrando de esta manera mayores acercamientos al manejo de la calidad en el proceso de planificación de consultas en SIBM.

Se planea extender el modelo de calidad para trabajar otros criterios diferentes, relacionados con la calidad del servicio, accesibilidad y seguridad, donde la intervención y definición de anotaciones por parte del usuario y del administrador del sistema, son un punto importante. Se planea además definir medidas para estos nuevos criterios, en la misma forma como se trabajó con la precisión y la cobertura.

Referencias bibliográficas

[1] José Luis Ambite y Craig A. Knoblock. Planning by rewriting: Eficiently generating high-quality plans. *Proceedings of the Fourteenth National Conference on Artificial Intelligence*, Providence, 1997.

[2] Donald Ballou, Stuart Madnick y Richard Wang. Special Section: Assuring Information Quality. *Journal of Management Information Systems*. Vol. 20, No. 3, 2004.

[3] Bennett, L., Wilkinson, G. y Oliver, K. The development and validation of instruments to assess the quality of Internet information: A progress report. *The Annual Convention of the Association for the Advancement of Computing in Education (AACE)*, Ed-Media 96, Boston Massachusetts. 1996.

[4] Manuel F. Bertoa, José M. Troya y Antonio Vallecillo. *Atributos de Calidad para Componentes COTS: Una valoración de la información ofrecida por los vendedores*. Di-sertación para Ph.D. Dpto. Lenguajes y Ciencias de la Computación. Universidad de Málaga. España, 2003.

[5] Ivan Catania, Martín Lobo, Pablo Miceli y Gonzalo Perez Bustelo. Metrica de Software. 2002. Fecha: 18-Mayo-05. Disponible en: http://members.lycos.co.uk/miguellobo/definiciones.htm.

[6] Ying Chen, Qiang Zhu, and Nengbin Wang. Query processing with quality control in the World Wide Web. *Journal World Wide Web*. Vol.1, Issue 4, 1998, pp. 241-255.

[7] Dvir, Ron y Evans Stephen . A TQM aproach to the improvement of Information Quality. *MIT Conference on Information Quality* Cranfield University, UK, 2000.

[8] Vladimir Estivill-Castro. Calidad total en informática. *Newsletters LANIA A.C.* Año 3, Vol 9, Otoño 1994. Disponible en: http://www.lania.mx/biblioteca/newsletters/1994-otono/art2.html. Fecha Acceso: 19-May-05.

[9] Beverly K. Kahn, Diane M. Strong y Richard Wang. Information Quality Benchmarks: Product and Service Performance. *Communications of the ACM*. Vol. 45 No. 4, 2002, pp.184-192.

[10] Craig A. Knoblock, Steven Minton, José Luis Ambite, Naveen Ashish, Ion Muslea, Andrew G. Philpot, y Sheila Tejada. The ARIADNE approach to Web-Based Information Integration. *International Journal of Cooperative Information Systems*. Originalmente publicado en AAAI'98. Vol. 10 No. 1-2, 2000, pp.145-169.

[11] Y. W. Lee, D. M. Strong, B. K. Kahn y R. Y. Wang. AIMQ: A Methodology for Information Quality Assessment. *Forthcoming in Information & Management*, Published by Elsevier Science. North Holland, 2001.

[12] Ulf Leser. Query Planning in Mediator Based Information Systems. PhD. Thesis Vom Fachbereich 13 –Informatik. Universitat Berlin, 2000.

[13] Ulf Leser y Felix Naumann. Query Planning with Information Quality Bounds. *Proceedings Flexible Query-Answering Systems*. 2000, pp.85-94.

[14] Alon Y. Levy, Alberto O. Mendelzon, Yehoshua Sagiv y Divesh Srivastava Answering queries using views. *Proceedings of the 14th ACM Symposium on Principles of Database Systems*, San Jose, California, 1995, pp.95-104.

[15] Alon Y. Levy, Anand Rajaraman y Joann J. Ordille. Query-Answering algorithms for information agents. *13th AAAI National Conference on Artificial Intelligence*, Publisher AAAI Press. Portland, Oregon, 1996, pp.40-47

[16] Alon Y. Levy. Logic-Based Techniques in Data Integration. *Kluwer International Series In Engineering And Computer Science*. Publisher Kluwer Academic Publishers,Seattle, WA. 2000, pp.575-595.

[17] Marotta, Adriana y Ruggia, Raul. Quality Management in Multi-Source Information Systems. *II Workshop de Bases de Datos*, Uruguay. 2002.

[18] Felix Naumann. Quality-driven Query Planning. Dissertation Outline Humboldt-Universitat zu Berlin, 2000.

[19] Felix Naumann. From Databases to Information Systems– Information Quality Makes the Difference. *Proceedings of the International Conference on Information Quality*. IBM Almaden Research Center, Cambridge, 2001.

[20] Felix Naumann, Ulf Leser and Johann Christoph Freytag. Quality-driven Integration of Heterogeneus Information Systems. *Proceedings of the International Conference on Very Large Databases (VLDB)*. Humboldt-Universitat zu Berlin, 2001, pp. 447-458.

[21] Naumann Felix y Rolker Claudia. Assessment Methods for Information Quality Criteria. *Proceedings of the International Conference on Information Quality (IQ)* Humboldt-Universität zu Berlin Forschungszentrum Informatik (FZI), Germany. 2000.

[22] OMG Unified Modeling Language Specification (draft). Version 1.3 beta R7. Object Management Group, The United States of America. June 1999. 798 p.

[23] Barbara Pernici. Data Quality evolution in Web Information Systems: model and management. *Proceedings. Lecture Notes in Computer Science 2503*. Springer 2002, pp. 397-413.

[24] Rachel Pottinger y Alon Levy. A Scalable Algorithm for Answering Queries Using Views. *Proceedings of the 26th VLDB Conference*. Cairo, Egypt, 2000, pp. 484-495.

[25] X. Qian. Query folding. *Proceedings of the Twelfth International Conference on Data Engineering ICDE*, New Orleans, LA, 1996, pp. 48–55.

[26] Risë L. Smith. Basic Research in the Virtual Library. For ENGL 101 and ENGL 201/301.Public Services Librarian & Associate Professor, Karl E.Mundt Library, Dakota State University. 1999.

[27] G. Wiederhold. Mediators in the architecture of future information systems. *IEEE Computer*, 1992, pp.38-49.

[28] Gene L. Wilkinson, Lisa T. Bennett, and Kevin M. Oliver. Evaluating the Quality of Internet Information Sources: Consolidated Listing of Evaluation Criteria and Quality Indicators. Department of Instructional Technology, University of Georgia, Athens, GA. Disponible en: http://it2.coe.uga.edu/Faculty/gwilkinson/webeval.html.

Bell Manrique Losada

Ingeniera de sistemas, Universidad Distrital Francisco José de Caldas en convenio con la Universidad de la Amazonia, Colombia. Magíster en ingeniería de sistemas, Universidad Nacional de Colombia, sede Medellín. Se desempeñó como docente de tiempo completo en la Universidad de la Amazonia donde lideró el grupo de investigación en Informática Educativa GIIE. Actualmente se desempeña como docente en la Universidad de Medellín y pertenece como investigadora al grupo de investigación ARKADIUS de la Universidad de Medellín. **bmanrique@udem.edu.co**

Jaime Alberto Guzmán Luna

Ingeniero civil, Universidad Nacional de Colombia, sede Medellín. Magíster en ingeniería de sistemas, Universidad Nacional de Colombia, sede Medellín. Actualmente realiza el Doctorado en Ingeniería en la Universidad Nacional de Colombia, sede Medellín. Se desempeñó como docente en la Universidad de Pamplona y en la Universidad Distrital. Actualmente se desempeña como profesor en la Universidad Nacional de Colombia, en Medellín y pertenece como investigador a los grupos GIDIA y SINTELWEB donde realiza estudios sobre Planificación en inteligencia artificial, web semántica y servicios web. **jaguzman@unal.edu.co**