Diccionario de Datos

Un enfoque semántico, de seguridad y usabilidad

Felipe McCalla, Iris González, Isabel Leguías, Xavier Medianero, estudiantes
Victor López , MSc.
Universidad Tecnológica de Panamá
Facultad de Ingeniería de Sistemas Computacionales
felipe.mccalla@utp.ac.pa, iris.gonzalez@utp.ac.pa, isabel.leguias@utp.ac.pa
xavier.medianero@utp.ac.pa, victor.lopez@utp.ac.pa

Resumen- *Este artículo trata sobre los diccionarios de datos enfocados en su utilidad semántica, léxica sobre las bases de datos y los esquemas relacionales; la manera en cómo debe ser estructurados y diseñados para su integración con los algoritmos ontológicos y con los sistemas de interfaces de usuarios. Además, el trabajo incluye la forma de extracción de datos implementando medidas de seguridad y a su vez medidas para la interpretación multilenguaje para la traducción del diccionario utilizando términos de base de datos. El artículo abstrae trabajos actuales sobre estas áreas, los integra y aporta nuevas ideas al tema.*

Palabras Claves - Diccionario de Datos, Multilenguaje, Ontología.

1. Introducción

Las bases de datos tienen como propósito el servir de repositorio para datos, mediante la implementación de modelos de DBMS (Sistema Manejador de Base de Datos) se pueden representar los datos por diferentes modelos, como por ejemplo: modelo relacional, jerárquico, de redes o de objetos, que ayudan a capturar la heterogeneidad de los datos. Los datos almacenados pueden ser observados por diferentes usuarios (usuarios de múltiples culturas) con diferentes niveles de abstracción, los cuales no manejan la misma terminología pudiendo causar malas interpretaciones de los mismos [1]. Además, los sistemas actuales son grandes y complejos, difíciles de controlar y manejar; para facilitar estas tareas se utilizan los denominadas "Diccionarios de Datos" los cuales presentan las siguientes ventajas:

a. Un diccionario de datos es un conjunto de metadatos que contienen las características lógicas y puntuales de los datos que se van a utilizar en el sistema, incluyendo descripción, alias, contenido y organización [2].
b. Los diccionarios se desarrollan durante el análisis de flujo de datos y ayudan a los analistas que participan en la determinación de los requerimientos del sistema, evitando así las ambigüedades, su contenido también se emplea durante el diseño del proyecto. Los elementos más importantes son los flujos de datos, almacenes de datos y procesos. El diccionario de datos guarda los detalles y descripción de todos estos elementos.

Los diccionarios de datos son utilizados porque permiten [3]:

a. Manejar los detalles en sistemas muy grandes, ya que tienen enormes cantidades de datos, aún en los sistemas más chicos hay gran cantidad de datos.
c. Facilitar el detalle de las características de las bases de datos relacionadas o del sistema en general y si son necesarias nuevas características.
d. Determinar dónde efectuar cambios en el sistema.
e. Localizar errores, omisiones en el sistema y detectar dificultades.
f. Aplicar un significado y terminología común para todos los elementos del sistema. Los diccionarios de datos proporcionan asistencia para asegurar significados comunes para los elementos y actividades del sistema.

Dada la importancia que presentan los diccionarios de datos en la representación y análisis de sistemas para base de datos grandes y complejos, su influencia sobre el flujo y tratamiento semántico de los datos, además de las características y aplicaciones mencionadas, se reconoce el impacto que tienen estos diccionarios.

En la aplicación de las mismas, es fundamental la integración de elementos críticos que permite aumentar la importancia y la utilidad de los diccionarios de datos (sobre todo para usuarios de culturas diferentes para la que fue hecha; esto se debe a que dada la expansión de la web se puede acceder desde cualquier lugar a la información.) Estos elementos corresponden a la utilización de ontologías, multilenguaje, técnicas eficientes de extracción de datos y mecanismos de seguridad, de tal manera es posible aumentar los beneficios que proporcionan, expandiéndolos a:

- Agilizamiento en procedimientos de extracción de dato [4].
- Mayor integridad e interoperabilidad con las bases de datos [5].
- Brindar la capacidad de soportar multilenguaje (capacidad de cambiar de idioma pero con un solo código) [6].
- Características ontológicas y semánticas para el significado de la información [7], [8].
- Características de protección de los datos del diccionario y a su vez de las bases de datos relacionadas.
- Diseño de Interfaces para la representación de los elementos de los diccionarios a los usuarios [9].

Este trabajo tiene como objetivo la integración de los elementos presentados, recopilando los beneficios, ventajas y facilidades de cada uno de ellos e integrándolos bajo un solo esquema; además de incorporar a cada componente características extras que permiten mejorar su utilidad en cada enfoque, para beneficio de una diversidad de usuarios siempre y cuando estos elementos presentados sean usables y amigables.

Este artículo está conformado de la siguiente forma: se explican los trabajos relacionados sobre las diferentes áreas a incorporar, posteriormente se integran aportando ideas para el mejoramiento de los diccionarios en cada área. Luego se discuten las ventajas enfocando la importancia de cada elemento.

2. Trabajos Relacionados

Debido a la importancia de la utilización de Diccionarios de Datos para facilitar el control de los datos tanto como su estructura, organización, significado semántico y seguridad de los mismos, se han elaborado diferentes trabajos, entre ellos los presentados a continuación:

2.1. Diccionario con Terminologías Multilenguaje

Lee y Park en [6], proponen métodos para remover la ambigüedad

en la interpretación de consultas en lenguaje natural a través de un *framework* denominado Gramática Categórica Combinatoria (CCG).

El método que utiliza CCG consiste en combinaciones extendidas de gramáticas categorizadas basadas en reglas de descomposición léxica y reducción en partes.

Presenta dos modelos de traducción: uno directo y uno indirecto, la representación en estos modelos no debe coincidir con objetos o elementos de la base de datos. La traducción basada en CCG deriva las expresiones base del lenguaje objetivo, ya sean instrucciones SQL, TSQL, lenguaje intermedio a partir del lenguaje base. El procedimiento de traducción utiliza una representación intermedia, lo cual proporciona diferentes ventajas, entre ellas: la independencia lingüística, la separación por dominios de conocimiento, la facilidad para determinar las inferencias y la reducción de la ambigüedad.

El *framework* CCG permite una traducción directa ente dos lenguajes, además de apoyarse en una representación intermedia. Las consultas que se realizan finalmente pueden ser formuladas en múltiples lenguajes proporcionándolo en el lenguaje base.

2.2. Interoperabilidad con Base de Datos

La interoperabilidad de los sistemas de base de datos está cobrando más importancia, debido a que las organizaciones pueden aumentar el número de sistemas operativos y añadir nuevos sistemas de apoyo a las decisiones [7],[5]. La construcción, operación y mantenimiento de estos sistemas es complicado, el tiempo y la complejidad crece más rápidamente si aumenta el número de sistemas [10].

El mayor problema de interoperabilidad consiste en combinar dos o más sistemas de bases de datos en un sistema coherente e integrado; ya que proporcionar la interoperabilidad entre los sistemas es mucho más difícil que la construcción de un sistema de base de datos distribuido.

En [5] se propone la integración automatizada por medio de la captura de datos utilizando la semántica de un diccionario normalizado. Esta arquitectura incorpora vistas locales definidas independientemente de la semántica de la base de datos, utilizando únicamente un diccionario predefinido entre sitios de integración. Mediante la normalización del diccionario en forma de cláusulas, se eliminan los conflictos de nombre, al igual que se reducen los conflictos semánticos.

La Base de Datos Semántica captura de forma independiente los documentos XML que son llamados X-Specs en los cuales se almacenan los nombres semánticos para los elementos de esquemas que determinan conceptos iguales en los sistemas. Luego se construye una vista integrada de conceptos que es transparente a las consultas que realiza el usuario. El procesador de consulta traduce las consultas semánticas a expresiones estructurales e integra los resultados.

2.3. Extracción Morfema usando Diccionarios

En [4], Nakamura y Yukishita investigan la extracción morfema de una cadena de fonemas por medio de las características de los diccionarios de datos. Para realizar la extracción de morfema, el diccionario de datos utiliza un método de acceso, una estructura de árbol de índice y una técnica en paralelo de índice. Los resultados obtenidos de estas técnicas son muy importantes para la prevención del cuello de botella en el procesamiento del lenguaje natural que se utilizan en la entrada de los datos, al igual que en la extracción de morfemas.

2.4. Interface para la Usabilidad de Diccionarios

En los trabajos presentados en [9] enfocan la importancia que presentan las interfaces dentro de los diccionarios de datos. Se establece un modelo basado en redes jerárquicas en el cual la información se relaciona por significado léxico con otra información del mismo diccionario, en donde se asocian todas las dependencias de un elemento padre con sus elementos hijos.

Cada red jerárquica contiene niveles de abstracción para cada uno de sus elementos. Un elemento identificado como tigre, se relaciona en niveles diferentes de transparencia con un elemento gato; el objetivo principal es la traducción de lenguajes determinando dinámicamente qué nodos y cuáles caminos conectan con la información que se desea trasladar.

2.5. Diccionarios Léxicos

La investigación en [8] presenta un sistema experto llamado lexicógrafico cuyo objetivo es de suplir al usuario con diferentes informaciones sobre las palabras rusas, incluyendo información bibliográfica concernientes a artículos del léxico individual. El sistema se concibe como una ayuda tanto en el ámbito del procesamiento del lenguaje natural y en la lexicografía tradicional, componiéndose así de dos componentes básicos: una base de datos bibliográfica y un lenguaje.

Las bases de datos léxicas son consideradas como un instrumento para predecir y calcular todo tipo de palabras de tipo semánticas. La base de datos léxica muestra una ventaja si la comparamos con un diccionario tradicional ya que una base de datos posibilita la presentación semántica de los dato utilizando un formato que facilita al computador arrojar una lista de palabras que poseen las mismas características comúnmente.

2.6. Seguridad en Base de Datos

Actualmente las base de datos han evolucionado en diferentes aspectos y esto incluye la seguridad. La seguridad en la base de datos [11] está basada en tres aspectos importantes: confidencialidad, integridad y disponibilidad. Además de estos aspectos hay otros que deben ser tomados en cuenta, como control de acceso, acceso a las aplicaciones, vulnerabilidades y los mecanismos de auditoría.

El método que generalmente se utiliza para brindar seguridad a los datos es restringir el acceso a los mismos, y se realiza a través del control de acceso, autenticación y autorización. Aunque estos mecanismos son diferentes, se utilizan en combinación con el control de acceso de granularidad para la asignación de derechos a objetos y usuarios. Dentro de una base de datos, estos objetos pueden ser tablas, vistas, filas y columnas. La limitación del acceso a los objetos se realiza a través de mecanismos de control de acceso Grant/Revoke. El control de acceso se especifica de tres maneras: Control de Acceso Obligatorio (MAC), Control de Acceso Discrecional (DAC) y Control de Acceso basado en Roles (RBAC). Tanto MAC como DAC proporcionan privilegios a los usuarios como a grupos asignados. Los roles son similares a las funciones

de trabajo. El objetivo principal de las funciones es la identificación de las operaciones y los objetos a los que estas operaciones necesitan tener acceso.

Políticas de Seguridad de base de datos

Las políticas de seguridad proporcionan una serie de directrices que soportan y orientan el proceso de seguridad en las bases de datos [12].

Las políticas sobre administración de seguridad en donde se destacan dos puntos importantes: el control centralizado, donde un solo administrador o grupo controla todos los aspectos de seguridad de la base de datos vs control descentralizado, diferentes administradores tienen control sobre diferentes partes de la BD, frecuentemente siguiendo lineamientos que se aplican a toda la BD. En el caso de propietario vs administrador, el propietario muchas veces se considera el responsable de los datos, pero cuando las BD son compartidas se requiere de un administrador cuyo objetivo es definir los datos compartidos por los usuarios y los controles de su uso.

Políticas para la especificación del control de acceso

Entre las cuales se encuentra las políticas del menor privilegio, máxima compartición de datos, sistemas abiertos y cerrados, control de acceso independiente del nombre, tipos de acceso, control independiente de la historia.

3. Discusión

Los diccionarios de datos tienen un papel relevante en el detalle y representación de los datos, detección de problemas y asociación de terminologías en relación a las bases de datos a las cuales están asociados. Su utilidad es incrementada cuando, en su diseño, se incorporan las siguientes características:

Seguridad: Debido a que en un diccionario de datos, los datos representan información del sistema que cubren, cualquier ataque que pueda revelar datos e información del mismo, proporcionará de forma indirecta información de cómo es el sistema, lo cual permitirá realizar un ataque más concreto y directo sobre las base de datos del sistema. De este hecho radica la importancia de asegurar la información y el acceso hacia los diccionarios de datos.

Usabilidad: Debido a que los datos presentes en los diccionarios de datos deben ser analizados y estudiados, las interfaces por las cuales se accede a esta información deben permitir el control completo de los datos, por lo que en este procedimiento deben aplicarse también los patrones de diseño de interfaces, para garantizar la usabilidad.

Además, es importante resaltar que los datos que se almacenan en los diccionarios de datos deben representar información relevante sobre las características e información extra que se necesite en el sistema.

Multilenguaje: La creación de diccionarios multilingües es muy valiosa ya que son una poderosa herramienta para almacenar datos. Mejor aún, si se contemplan varios idiomas para el mismo diccionario se podría entonces utilizar por una diversidad de usuario. Para lograr una producción de calidad en el diccionario multilenguaje es propicio integrar las herramientas informáticas dominado por lexicógrafos, que son esencialmente los procesadores de palabras.

Semántica: Los elementos almacenados deben otorgar un valor semántico al diccionario de datos como se muestra en la sección de usabilidad. De esta forma mediante las interfaces no sólo se buscarán elementos puntuales (léxicos), sino que facilitará la búsqueda ontológica de elementos (semántica) lo cual permitirá al usuario la búsqueda de otras características.

Interoperabilidad: Otro elemento a considerar dentro de los diccionarios, es la interoperabilidad que existe entre el mismo y las bases de datos. Por eso en [5] se menciona que la integración de fuentes de datos implica la combinación de sus conceptos y el conocimiento en una visión integrada que aísla a los usuarios de la organización del sistema. Sin embargo, en [7] señalan los criterios y técnicas a seguir para apoyar el establecimiento de un diccionario semántico.

Las características presentadas, aumentan la utilidad que presentan los diccionarios de datos e incorporan ventajas extras a la aplicación de diccionarios dentro de un sistema.

El tratamiento y la traducción de frases o sentencias en lenguaje natural a otros lenguajes, es una de las tareas actuales que deben ser incorporadas dentro de los sistemas debido al fenómeno moderno de internalización de la información.

Procedimientos de descomposición léxica del idioma base en elementos indicadores como sujetos, predicados y verbos, en conjunto con el valor semántico que pueden agregar los mismos permite que el proceso de traducción sea más eficiente pues se identifican las capas de las mismas.

Además, es necesario manifestar que en las Interfaces de usuarios no sólo es importante la información que se va a presentar y la usabilidad de la misma, sino que también es necesario aplicar patrones de multiculturalidad tanto en las Interfaces, como al momento de manipulación de los datos del diccionario, debido a que el significado de los mismos puede variar el sentido real causando inconsistencia.

4. Comentarios Finales

Cuando un sistema amerita la integración de Diccionarios de Datos, no sólo se deben enfocar los aspectos y ventajas propias del mismo como el detalle de las características del sistema, localización de errores, entre otros; sino también incluir aspectos de seguridad para protección de datos de forma directa e indirecta, es decir, proteger los datos del diccionario y cualquier información que se pueda obtener del sistema a través de la visualización de los mismos, aspectos de multilenguaje orientados a multiculturalidad, de tal forma que la información recabada en los diccionarios sea más útil para los actores que utilizan la misma, sin olvidar que las interfaces de consulta a los mismos también deben integrar esos patrones.

Además, la efectividad de la utilización de los diccionarios, puede ser aumentada incorporando técnicas semánticas y ontológicas, pues permite no sólo captar datos aislados, sino comprender el objetivo de los mismos, por lo que también añadirá ventajas en búsqueda y consultas.

Referencias

[1] C. Bartini, et al., Diseño Conceptual de Base de Datos. Massachusetts, United States Addison-Wesley Iberoamericana, S.A., 1994.

[2] A. Silberschatz, et al., Fundamentos de Base de Datos, 4 ed., 2001.

[3] M. P. Meza. (Consultado: 5 de octubre de 2010, Tutorial de Administración de Base de Datos. Available: http://sistemas.itlp.edu.mx/tutoriales/admonbasedat/tema5_1.htm

[4] O. Nakamura and M. Yikishita, "A High-speed Morpheme-Extraction System Using Dictionary Database," in Proceedings. Fourth International Conference Data Engineering, Los Angeles, CA, 1988, pp. 488-495.

[5] R. Lawrences and K. Baker, "Integrating Relational Database Schemes using a Standardized Dictionary," in Simposio sobre Informática Aplicada Actas del simposio de ACM New York, EE.UU., 2001, pp. 225-230.

[6] H. Lee and J. Parque, "Automatic Augmentation of Translation Dictionary with Database Terminologies in Multilingual Query Interpretation," in Actas del taller sobre Tecnología del Lenguaje Humano y Gestión del Conocimiento, Toulouse, Francia, 2001.

[7] S. Castano and V. D. Antonellis, "Semantic Dictionary Design for Database Interoperability," in Data Engineering, Proceedings, 13th International Conference Birmingham, UK, 1997, pp. 43-54.

[8] E. Paducheva, et al., "Semantic dictionary viewed as a lexical database," in Proceedings of the 14th conference on Computational linguistics, Actas de COL1NG-92, Nantes, Francia, 1992, pp. 1294-1298.

[9] H. Ozawa, et al., "DIS: A User Interface System design for the dictionary's database," Languages for Automation: Symbiotic and Intelligent Robots, IEEE Workshop pp. 164-169, 1988.

[10]R. Lawrence, "Automatic Conflict Resolution to Integrate Relational Schema," Ph,D, Doctor of Philosophy in Computer Science, UNIVERSITY of MANITOBA, Canada, 2001.

[11] M. Coffin, "Database Security: What Students Need to Know," Journal of Information Technology Education, vol. 9-2010, 2010.

[12] M. López, "Diccionario/Directorio y Seguridad de Datos," Centro de investigación en Sistemas de Información.