Descubriendo Conocimiento a partir de las Dependencias de Comparación de Conjuntos de Valores

Viviana E. Ferraggine [1] Laura C. Rivero [1, 2]

[1] INTIA, Facultad de Ciencias Exactas, U.N.C.P.B.A., Tandil, Buenos Aires, Argentina
[2] LINTI, Facultad de Informática, U.N.L.P. La Plata, Buenos Aires, Argentina
{vferra, lrivero} at exa.unicen.edu.ar

RESUMEN

El desarrollo y la investigación sobre modelos de datos semánticos se han visto estimulados por los requerimientos de mayor expresividad en modelos de datos conceptuales. Por esto, la identificación y representación de reglas que son de sencilla interpretación en el mundo real pero de difícil representación en los modelos conceptuales de datos han merecido un estudio particular.

Cuando en el Universo de Discurso (UdeD) existen conjuntos de atributos u objetos compatibles y semánticamente vinculados se presentan frecuentes comparaciones entre ellos. En los modelos de datos, esto da lugar a una serie de construcciones y de restricciones que pueden representarse por medio de *Dependencias de Comparación de Conjuntos de Valores (**dccv**)* como limitantes del modo en el cual un conjunto de datos se relaciona con otro. Pueden establecerse cuatro formas de comparación: *inclusión*, *exclusión*, *igualdad*, *superposición parcial* o *solapamiento*. Este trabajo está dedicado a plantear un proceso de análisis de dichas dependencias que va desde realizar la *correspondencia de la dependencia con una estructura de modelado* hasta llegar al *análisis de la estructura lingüística* (en lenguaje natural), y previamente habiendo realizado *el análisis de su origen,* con el objetivo de contribuir a la elicitación de conocimiento del Universo del Discurso (UdD).

Palabras clave: Diseño conceptual, dependencias de comparación de conjuntos de valores, dependencias de inclusión, dependencias de igualdad, dependencias de exclusión, dependencias de superposición parcial, restricciones, modelos de datos semánticos y conceptuales.

CONTEXTO

Este trabajo forma parte de las actividades científicas de la Línea "Integridad en Base de Datos guiada por Requisitos", del Proyecto "Base de Datos y Procesamiento de Señales". Este proyecto es continuación de proyectos previos en la temática y está actualmente en desarrollo en el Instituto de Investigación en Tecnología Informática Avanzada (INTIA) de la Facultad de Ciencias Exactas de la U.N.C.P.B.A.

El objetivo general de la línea es estudiar los problemas de la elicitación de conocimiento del UdeD, así como el impacto de la realidad estudiada sobre los problemas de modelado e integridad en el contexto de bases de datos relacionales y post-relacionales.

1. INTRODUCCIÓN

La adquisición de conocimientos del dominio es crucial para la calidad de la base de datos de destino y el estudio de las dependencias que existen entre los datos del mundo real ofrece herramientas muy valiosas para la adquisición de conocimiento semántico en la fase conceptual del diseño de base de datos.

Las restricciones de integridad son condiciones que capturan la semántica del dominio de aplicación en cuestión pero en la práctica, la decisión de especificar una restricción es muy importante y muy difícil.

Existen dos dificultades principales cuando se trata de especificar las propiedades de los datos dentro del contexto de modelado conceptual de datos. Primero son difíciles de categorizar, en especial las que son evidentes en el mundo real y segundo, su representación varía durante el proceso de diseño de software.

El primer aspecto cae en el campo de la Ingeniería de Requisitos [12] mientras que el

restante cae en el campo del Modelado Conceptual y Modelado Lógico Temprano o genérico. Durante estas actividades, las particularidades del modelo de datos generan una diversidad de formas en las que estos aspectos pueden ser considerados.

En el mundo real, las propiedades de los datos se perciben desde un punto de vista semántico, permitiendo su agrupación según la complejidad semántica intrínseca y su clasificación de acuerdo a su esencia. Pero en el mundo de las bases de datos se perciben básicamente desde un punto de vista sintáctico. A medida que el desarrollo del artefacto del software avanza, deben encontrarse representaciones de estas propiedades según una estructura lícita del modelo. Esta representación será ciertamente diferente en el modelo de datos conceptual, en el modelo de datos lógico y en el modelo de datos del motor de la base de datos [13, 14, 15].

Aunque los modelos conceptuales actuales han contribuido en gran medida a la captura de los aspectos relevantes del UdeD, en algunos casos el poder expresivo no es suficiente para representar integralmente la diversidad natural de reglas del negocio. Los medios formales o semiformales que éstos ofrecen permiten representar parte de su semántica, sobre todo cuando se trata de reglas estáticas. Muchas de ellas pueden ser representadas en algunas de las variantes del Modelo de Entidades y Relaciones (MER) [6, 17], por ejemplo por medio de restricciones de cardinalidad y dependencias de existencia. Otras reglas que requieren la formulación explícita de eventos, condiciones y acciones específicos, no pueden ser representadas de un modo equivalente. Esto ha promovido la materialización de numerosas propuestas de ampliación a este modelo [2, 3, 4, 10, 11]. Por otro lado, algunas restricciones de integridad que podrían haberse representado adecuadamente con los mecanismos de abstracción provistos por el modelo, se difieren infundadamente hasta etapas más tardías del proceso de diseño y frecuentemente su materialización no utiliza los recursos expresivos más apropiados para el caso.

Desde otro punto de vista, existen enfoques en defensa del MER original proponiendo una rigurosa y amplia utilización de las capacidades que éste ofrece, sin las extensiones que dificultarían su uso, interpretación, validación y formalización [9]. Adhiriendo a este último enfoque, este trabajo se desarrolla en dos direcciones complementarias: 1) ofreciendo siempre que sea posible un modo de representar cada regla utilizando las construcciones básicas del MER y cuando esto no sucede proveyendo patrones de lenguaje declarativo; 2) sugiriendo estrategias de transformación en función de lo obtenido en 1), cuando una regla se ha plasmado de forma incorrecta o ambigua.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Una de las líneas incluidas en nuestras actividades de investigación está centrada actualmente en un tipo de reglas de sencilla interpretación en el mundo real pero con dificultades manifiestas de representación en el MER: las "*Dependencias de Comparación de Conjuntos de Valores*" (***dccv***), nombre genérico otorgado a las expresiones que permiten establecer comparaciones entre los valores de conjuntos de atributos estructuralmente compatibles y semánticamente relacionados. Esta clase de reglas incluye dependencias de inclusión (***di***) [5, 7]; dependencias de exclusión (***de***) [1, 8]; dependencias de igualdad (***dig***) y dependencias de superposición parcial (***dsp***)[16].

La propuesta sugiere analizar las ***dccv*** desde una base sintáctica, a fin de establecer: a) la posibilidad de su representación en un MER utilizando las estructuras estandarizadas; b) la necesidad de ampliar el modelo para capturar la semántica de esas reglas, cuando las estructuras estandarizadas no sean suficientes, y c) la posibilidad de realizar transformaciones estructurales que podrían proporcionar algún beneficio en relación con la semántica o la calidad de los aspectos operativos, d) otros medios de representación

cuando los medios analíticos disponibles sean insuficientes, [13, 14, 15], e) las bases de un análisis de la estructura lingüística que dio origen a dicha regla.

1.1. Aspectos del Proceso de Análisis de las *dccv*:

El proceso de análisis de las ***dccv*** se inicia con el *análisis de su origen* para posteriormente realizar la *correspondencia de la dependencia con una estructura de modelado* hasta llegar al *análisis de la estructura lingüística* (en lenguaje natural) de la misma.

En primer lugar se realiza una *descripción de la dependencia* identificando los dos términos que la componen; se los denomina término del lado izquierdo y término del lado derecho de la dependencia, en correspondencia a la denominación que reciben los mismos cuando corresponden a dependencias particulares como son las restricciones de integridad referencial. Adicionalmente se inspecciona la conformación de ambos términos para identificar y clasificar el conjunto de atributos que lo componen. Para lo anterior se discrimina si el término está compuesto por atributos que forman la clave de un objeto (o lo identifican) o parte de ella o si son simplemente atributos descriptores del objeto.

Si existe alguna posible *transformación* del esquema en función de las dependencias funcionales o de inclusión planteada es posible aplicar algunos axiomas de transformación básicos del modelo relacional, como son la *regla de pullback* y el *axioma de transitividad*. Existen otros casos en los que transformaciones sencillas basadas en atributos y relaciones inaplicables [7] mejoran la calidad respecto a su representación. También hay un conjunto de casos en los que la calidad semántica no mejora dado que las transformaciones no permiten revelar objetos y relaciones mal representados, o bien porque no hay transformación que sea aplicable.

Para *el análisis del origen de la dependencia* se inspecciona la composición de cada término tratando de identificar su posible origen en el contexto de un modelo lógico o más específicamente de un modelo relacional, para luego hacer una ingeniería reversa e identificar los objetos y los atributos que intervienen.

Luego, para realizar el *análisis correspondiente a la estructura del modelo conceptual* se transforma el modelo lógico en su o sus correspondientes alternativas de un modelo de entidades y relaciones extendido (MERE) lo que posibilita estudiar con que semántica se ha plasmado dicha dependencia.

Por último para llegar a *analizar la estructura lingüística de la dependencia* se transforman las construcciones del MERE en oraciones simples donde las entidades se transforman en sustantivos que funcionan como sujeto y objeto directo y las relaciones son el verbo que realiza el nexo.

1.2. Aspectos Formales

Los aspectos formales que se han tomado en cuenta para realizar el análisis del origen de la dependencia parten de realizar un patrón de clasificación para cada una de ellas.

La estructura sintáctica de los términos de una dependencia puede definirse en función de la posición del término respecto de la clave de la relación (correlación). Siendo W un conjunto de atributos de una cierta relación R, K la clave primaria de R y Z un subconjunto de atributos no pertenecientes a la clave, existen cinco posiciones posibles para los términos. I) $W \equiv K$; II) $W \equiv Z$; III) $W \equiv K^1$; IV) $W \equiv K \cup Z$; V) $W \equiv K^1 \cup Z$ (K^1 es un subconjunto estricto de K, $K^1 \neq \emptyset$ y $Z \neq \emptyset$ para todos los casos).

Sean R_i y R_d relaciones, denominadas término izquierdo y derecho respectivamente; W_i y W_d conjuntos de atributos compatibles de R_i y R_d. Teniendo en cuenta las correlaciones se tienen 25 posibles casos de ***dccv*** $<R_i.W_i\ \theta\ R_d.W_d>$ que son los siguientes:

$1:<K_i\ \theta\ K_d>$; $2:<Z_i\ \theta\ K_d>$; $3:<K_i^1\ \theta\ K_d>$; $4:<K_i \cup Z_i\ \theta\ K_d>$; $5:<K_i^1 \cup Z_i\ \theta\ K_d>$; $6:<K_i\ \theta\ Z_d>$; $7:<Z_i\ \theta\ Z_d>$; $8:<K_i^1\ \theta\ Z_d>$; $9:<K_i \cup Z_i\ \theta\ Z_d>$; $10:<K_i^1 \cup Z_i\ \theta\ Z_d>$; $11:<K_i\ \theta\ K_d^1>$; $12:<Z_i\ \theta\ K_d^1>$; $13:<K_i^1\ \theta\ K_d^1>$; $14:<K_i \cup Z_i\ \theta\ K_d^1>$; $15:<K_i^1 \cup Z_i\ \theta\ K_d^1>$; $16:<K_i\ \theta\ K_d \cup Z_d>$; $17:<Z_i\ \theta\ K_d \cup Z_d>$; $18:<K_i^1\ \theta\ K_d \cup Z_d>$; $19:<K_i \cup Z_i\ \theta\ K_d \cup Z_d>$; $20:<K_i^1 \cup Z_i\ \theta\ K_d \cup Z_d>$; $21:<K_i\ \theta\ K_d^1 \cup Z_d>$; $22:<Z_i\ \theta\ K_d^1 \cup Z_d>$;

23:<K_i^1 θ $K_d^1 \cup Z_d$>; 24:<$K_i \cup Z_i$ θ $K_d^1 \cup Z_d$>; 25:<$K_i^1 \cup Z_i$ θ $K_d^1 \cup Z_d$>

Se define cada uno de los cuatro tipos de ***dccv*** de la siguiente manera:

Una *dependencia de inclusión* (***di***) se define como la existencia de atributos en una relación cuyos valores deben ser un subconjunto de los valores de atributos compatibles en otra relación (o la misma). Formalmente una ***di*** es una expresión $\mathbf{R_i[W_i] \subseteq R_d[W_d]}$. Si W_d es la clave primaria (K_d) de R_d, la ***di*** es basada en clave y se la denomina restricción de integridad referencial (***rir***) o simplemente referencia.

Una *dependencia de igualdad* (***dig***) se define como la existencia de un conjunto de atributos en una relación cuyos valores deben ser los mismos que los correspondientes a un conjunto de atributos compatibles en otra relación (o la misma). Esto significa que cada miembro del primer conjunto debe ser un miembro del segundo conjunto y viceversa.
Formalmente una ***dig*** se indica $\mathbf{R_i[W_i] = R_d[W_d]}$

Una *dependencia de exclusión* (***de***) se define como la existencia de un conjunto de atributos en una relación cuyos valores deben ser mutuamente excluyentes respecto de los correspondientes a un conjunto de atributos compatibles en otra relación (o en la misma). Esto significa que ambos conjuntos de atributos no tienen miembros en común.
Formalmente una ***de*** es una expresión $\mathbf{R_i[W_i] \,||\, R_d[W_d]}$

Una *dependencia de superposición parcial* (***dsp***) se define como la existencia de un conjunto de atributos en una relación cuyos valores se solapan parcialmente con los correspondientes a un conjunto de atributos compatibles en otra relación (o en la misma).
Formalmente una (***dps***) se denota $\mathbf{R_i[W_i] \pitchfork R_d[W_d]}$ y tiene un significado equivalente a: $R_d.W_d \not\subseteq R_i.W_i \wedge R_i.W_i \not\subseteq R_d.W_d \wedge R_d.W_d \pitchfork R_i.W_i$.

Para el análisis de las ***dccv***s se ha seguido lo expuesto en [14, 15], examinando la estructura sintáctica de los conjuntos de atributos involucrados, independientemente del dominio de aplicación. Esto permite inferir el posible tipo de vínculo semántico que intenta plasmar la dependencia. Mediante la heurística descripta allí es posible realizar un rediseño parcial del esquema y sus restricciones, y esto mejorará su calidad con referencia a sus aspectos semánticos.

1.3. Análisis de Casos.

A continuación se aplica el proceso de análisis de un caso de ***dccv.*** Sean las siguientes relaciones, con sus claves subrayadas y las ***rir***s correspondientes:

R_i: PROFESOR (id_P, id_D, NombreP)
R_d: CURSO (id_C, id_P , id_D, Descripción)
DEPARTAMENTO (id_D, NombreD)
rir1: CURSO.id_P << PROFESOR.id_P
rir2: CURSO.id_D << DEPARTAMENTO.id_D
rir3: PROFESOR.id_D<< DEPARTAMENTO.id_D
Ningún atributo admite nulos y se satisface la ***dccv*** CURSO (id_P, id_D) ⊆ PROFESOR(id_P, id_D) que indica la *regla del negocio* "los cursos de un departamento ***deben*** ser dictados por profesores que ***deben*** pertenecer al mismo departamento".

Descripción de la dependencia: el término izquierdo de la dependencia es un conjunto de atributos que no conforman la clave del objeto izquierdo y el término derecho es la unión de la clave con un conjunto de atributos secundarios, tipificado como caso 17:<Z_i θ $K_d \cup Z_d$> de la ***dccv,*** donde θ es una ***di*** (⊆)

Transformación (si se requiere):

Dado que además de la ***di*** CURSO (id_P, id_D) ⊆ PROFESOR (id_P, id_D) existe una dependencia funcional
PROFESOR.id_P → PROFESOR.id_D.
puede aplicarse la regla de "pullback" por la cual se puede inferir CURSO.id_P → CURSO.id_D. que hace redundante la *rir2.*
Por lo anterior el esquema transformado es:
R_i: PROFESOR (id_P, id_D, NombreP)
R_d: CURSO (id_C, id_P , Descripción)
DEPARTAMENTO (id_D, NombreD)
rir1: CURSO.id_P << PROFESOR.id_P
rir3: PROFESOR.id_D<< DEPARTAMENTO.id_D

Correspondencia de la dependencia con una estructura de modelado: para éste caso puede expresarse que la correspondencia después de la transformación ha convertido una ***di pura*** en otra basada en clave (***rir***), catalogada como ***dccv*** correspondiente al grupo 2:<Z_i θ K_d>. El patrón de modelado que corresponde a una dependencia de inclusión entre conjuntos de valores que corresponden a atributos que son claves en el objeto derecho y atributos secundarios en el objeto izquierdo es el de una relación de cardinalidad máxima N:1.

Análisis de la estructura lingüística de la dependencia: la estructura lingüística de la frase que describiría este aspecto en el UdD dado que todos los atributos son obligatorios, podría ser "Cada curso **debe** ser dictado por un **profesor y el profesor debe** pertenecer a un departamento" u otra equivalente. Y la aplicación de los axiomas de Armstrong a las dependencias funcionales permite completar esta interpretación deduciendo que el departamento de profesor y curso es el mismo.

Se observa que para estos casos de ***dccv*** la estructura que se percibe es de dos sentencias ambas compuestas por sujeto, verbo y objeto directo, a diferencia de una dependencia funcional que daría lugar solo a una sentencia. De acuerdo a los verbos utilizados es posible identificarlos de acuerdo a dos tipos: los que manifiestan obligatoriedad (por ej. *puede* tener, ser, pertenecer, poseer, etc.) y los que manifiestan opcionalidad (*debe* tener, ser, pertenecer, poseer, etc.).

3. RESULTADOS OBTENIDOS Y ESPERADOS

Se han determinado los patrones de representación de ***di***s y ***dig***s en función de ***rir***s. En base a los logros anteriores, se han evaluado las ventajas y desventajas de aplicar las transformaciones según el nivel semántico del esquema original y la implementación de las operaciones de actualización básicas. Se pueden resumir las siguientes conclusiones parciales:

- Siempre que sea conveniente la transformación, el nuevo esquema adhiere al MER convencional o al MER extendido con agregaciones.
- Con respecto a razones de diseño, los esquemas transformados generalmente introducen nuevas tablas. Esto produce un nuevo contexto de desempeño que debería ser analizado teniendo en cuenta las características del artefacto de software.
- Finalmente, el esquema resultante debe re-analizarse con base en las expresiones de multiplicidades asociadas a los roles de las entidades participantes para determinar la validez de los caminos cíclicos y eliminar relaciones redundantes. Debido al agregado de tablas y relaciones se hace necesario un nuevo análisis de la coherencia global del esquema resultante.

La aplicación de transformaciones (en conjunción con el análisis de vínculos semánticos redundantes y el uso de mecanismos de evaluación de esquemas) ha resultado un marco metodológico apropiado para lograr una parcial pero útil reingeniería de esquemas de base de datos relacionales de pobre calidad, desactualizados o erróneos. La fase de transformación contribuye en el mejoramiento del nivel semántico del esquema, poniendo en evidencia objetos mal modelados y relaciones que los unen con otros objetos. Las transformaciones aplicables a las dependencias de inclusión y concretadas a través de una heurística fueron apropiadas también para las dependencias de igualdad. Algunos casos no pudieron mejorarse con este enfoque debido a que persisten dependencias intratables. No es posible establecer una regla estricta a seguir, pero generalmente, si una relación es conceptualmente relevante o las operaciones de actualización son poco frecuentes, la transformación es conveniente aunque aumentará el número de entidades y relaciones.

El estudio sobre las ***de***s y ***dsp***s ha sido enfocada desde otra perspectiva ya que ninguna de ambas está en el espíritu del MER, con la excepción de las jerarquías exclusivas y compartidas respectivamente.

Dentro de los futuros trabajos de investigación se incluye la confección de una plantilla lingüística para cada uno de los casos y el análisis de las acciones que podrían desencadenar las ***dccv***s haciendo una analogía con las acciones referenciales de las ***rir***s y el estudio de patrones de código (triggers) para soportar las ***di***s remanentes en cada caso.

4. FORMACION DE RECURSOS HUMANOS

La investigación desarrollada ha permitido elaborar un programa para materias optativas centradas en tópicos avanzados de modelado conceptual de datos. En el mismo sentido varios grupos de alumnos han desarrollado o han profundizado en aspectos conceptuales relativos a este tema durante el desarrollo de su tesis de grado para aspirar al título de Ingeniero de Sistemas-UNCPBA.

5. BIBLIOGRAFÍA

[1] Albrecht, M., Buchholz, E., Düsterhöft, A., Thalheim, B., "An Informal and Efficient Approach for Obtaining Semantic Constraints using Sample Data and Natural Language Processing", Lecture Notes In Computer Science, 1358, pp. 1-28. Selected Papers from a 1st. Workshop on Semantics in Databases, Czech Republic. Thalheim, B.&Libkin, L.(Eds.), 1998.

[2] Badía, A., "Entity-Relationship Modeling Revisited", SIGMOD Record, 33(1), 2004.

[3] Camps, R., "From Ternary Relationship to Relational Tables: a Case Against Common Beliefs", SIGMOD Record, 31(2), pp. 46-49, 2002.

[4] Camps, R., "Transforming N-Ary Relationships to Database Schemas: an Old and Forgotten Problem", RR LSI-02-5-R, Univ.Politécnica de Catalunya, Spain, 2002.

[5] Casanova, M., Fagin, R., Papadimitriou, C., "Inclusion Dependencies and their Interaction with Functional Dependencies", J. Computer and System

Sciences, 28(1), 1984.
[6] Chen, P., "The Entity-relationship Model: Toward a Unified View of Data", ACM TODS, 1(1), pp. 9-36, 1976.
[7] Codd, E., "The Relational Model for Database Management. Version 2", Addison Wesley Publ. Co., 1990.
[8] De Miguel, A., Piattini, M., Marcos, E., "Diseño de Bases de Datos Relacionales", Alfaomega Grupo Editor, 2000.
[9] Goelman, D., Song, I-Y., "Entity-Relationship Modeling Re-revisited", ER 2004, Proceedings 23rd International Conference on Conceptual Modeling, Shanghai, China. LNCS, Springer, 3288, Atzeni P., Wesley W. Chu, Hongjun Lu, Shuigeng Zhou, Tok Wang Ling Editors, pp. 43-54, 2004.
[10] Halpin, T., "Information Modeling and Relational Databases. From Conceptual Analysis to Logical Design", Morgan Kaufmann Publishers, 2001.
[11] Jones, T., Song, I-Y., "Analysis of Binary/Ternary Cardinality Combinations in Entity-Relationship Modeling", Data & Knowledge Engineering, 19 (1), pp. 39-64, 1996.
[12] Loucopoulos, P & Karakostas. (1995). System Requirements Engineering. McGraw Hill International Series in Software Engineering.
[13] Rivero, L., "Inclusion Dependencies", In Encyclopedia of Information Science and Technology. Idea Group. Inc., Mehdi Khosrow-Pour Editor, pp. 1425-1430, 2005.
[14] Rivero, L., Doorn, J., Ferraggine, V., "Inclusion Dependencies". In Developing Quality Complex Database Systems: Practices, Techniques and Technologies, Idea Group Inc., Shirley A. Becker Editor, pp 261-278, 2001.
[15] Rivero, L., Doorn, J., Ferraggine, V., "Enhancing Relational Schemas through the Analysis of Inclusion Dependencies", Int'l Journal of Computer Research, 12(4), Nova Publishers, pp. 489-511, 2004.
[16] Tari, Z., Bukhres O., Stokes J., Hammoudi S., "The Reengineering of Relational Databases Based on Key and Data Correlations", In Searching for Semantics: Datamining, Reverse Engineering, etc. Proc.7th IFIP 2.6 Working Conference on Database Semantics (DS-7), Chapman & Hall, Lausanne, S. Scappapietra and F. Maryanski Editors, pp. 183-214, 1998.
[17] Teorey, T.J., "Database Modeling and Design. The Entity-Relationship Approach", Morgan Kaufmann Publishers, 1990.