Análisis de rendimiento académico estudiantil usando Data Warehouse Difuso

Analysis of students' academic performance using Fuzzy Data Warehouse

Carolina Zambrano Matamala[1]* Angélica Urrutia Sepúlveda[2] Marcela Varas Contreras[3]

Recibido 7 de septiembre de 2015, aceptado 18 de julio de 2016
Received: September 7, 2015 Accepted: July 18, 2016

RESUMEN

Un Data Warehouse (DW) es un repositorio de datos que provienen de distintas fuentes. Es utilizado para el análisis de datos y como apoyo a la toma de decisiones. Ya que permite obtener indicadores cuantitativos tales como cantidades. Sin embargo, en el ámbito de análisis de datos es usual encontrar relaciones entre los datos de naturaleza difusa. Por ejemplo, en un contexto académico la respuesta a la consulta "qué estudiantes obtuvieron buena nota final" no puede ser obtenida desde un DW tradicional pues no maneja información cualitativa. Debido a las limitaciones de los DW tradicionales es que pueden extenderse usando lógica difusa a un DW Difuso (DWD). En este trabajo se presenta una implementación de DWD cuyo objetivo es permitir el análisis cualitativo de datos de estudiantes lo que apoya el proceso de toma de decisiones. El DWD implementado permite operar con medidas difusas, hechos difusos, relaciones difusas y niveles difusos.

Palabras clave: Data Warehouse Difuso, lógica difusa, Data Warehouse.

ABSTRACT

A Data Warehouse (DW) is a data repository containing data from different sources. It is used to support decision-making through data analysis as it allows quantitative indicators such as amounts. However, data analysis lack of data of diffuse nature. For instance, in an academic context, the answer to the query "What students scored with good final grade" cannot be obtained from a traditional DW because they do not handle qualitative information. Due to the limitations of traditional DW the use of fuzzy logic may extend its capabilities through a fuzzy DW (FDW). This paper presents an implementation of FDW aimed at allowing qualitative analysis of scholar data supporting decision making process. The FDW implemented operate fuzzy measures, fuzzy facts, fuzzy relations and fuzzy levels.

Keywords: Fuzzy Data Warehouse, fuzzy logic, Data Warehouse.

INTRODUCCIÓN

Cada día las organizaciones tienen más información porque sus sistemas producen una gran cantidad de operaciones diarias que se almacenan generalmente en bases de datos transaccionales. Con el fin de analizar esta información histórica, una alternativa interesante es implementar un DW. Un DW es una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado,

[1] Escuela de Ingenierías. Universidad Católica del Norte. Larrondo 1281, Coquimbo, Chile. E-mail: carolinazambrano@gmail.com
[2] Departamento de Computación e Informática. Universidad Católica del Maule. Avenida San Miguel 3605. Talca, Chile. E-mail: aurrutia@spock.ucm.cl
[3] Departamento de Ingeniería Informática y Ciencias de la Computación. Universidad de Concepción. Edmundo Larenas 215. Concepción, Chile. E-mail: mvaras@udec.cl
* Autor de correspondencia

no volátil y variable en el tiempo, que es utilizado para el apoyo a la toma de decisiones en la entidad en la que se utiliza [3, 12].

Los elementos principales de un DW son dimensiones, jerarquías, hechos y medidas.

Una dimensión es un concepto abstracto que modela un contexto para el análisis. Sobre las dimensiones se pueden definir jerarquías, que permiten acceder a los datos a diferentes niveles o categorías de detalle [3, 12].

Los hechos representan una tupla de claves primarias, foráneas y medidas. Las medidas son valores numéricos, por lo que el tipo de análisis que se puede realizar es cuantitativo. Sin embargo, en el proceso de toma de decisiones también es valorado el análisis cualitativo. El análisis cualitativo, está orientado a analizar características de alguna cosa más que una cantidad que es el caso de un análisis cuantitativo.

En el contexto de análisis cuantitativo se utilizan los DW y para extender un DW a un DWD se utiliza lógica difusa [25] de tal forma de incluir los aspectos cualitativos en el análisis de datos. Entonces es necesario comprender cuáles son los elementos de un DW que se extienden para componer un DWD. Comenzando por qué es un atributo difuso. Una clasificación de información difusa es presentada en [9-10, 22, 29-30] donde se definen los atributos difusos como atributos tipo 1, tipo 2 y tipo 3.

Los atributos difusos tipo 1 [9-10, 22, 29-30] son atributos clásicos que admiten el tratamiento impreciso, donde las etiquetas lingüísticas definidas solo se usarán en las condiciones difusas de las consultas.

Los atributos difusos tipo 2 [9-10, 22, 29-30] son atributos que admiten tanto datos clásicos (crisp) como difusos (imprecisos), en forma de distribuciones de posibilidad sobre un dominio subyacente ordenado. El tipo 2 permite también la representación de información incompleta en forma de datos tipo Unknown, Undefined y Null.

Los atributos difusos tipo 3 [9-10, 22, 29-30] son atributos sobre datos de dominio discreto sobre dominio subyacente no ordenado con analogía. En estos atributos se pueden definir conjuntos de etiquetas con una relación de similitud o proximidad definida sobre ellas. Además, en el tipo 3 se pueden admitir distribuciones de posibilidad sobre el dominio.

Por otro lado, la mayoría de los artículos y casos de estudio que incluyen análisis de datos usando DWD están relacionados con el impacto o beneficio de aplicar DWD a contextos empresariales [6, 15]. Sin considerar el aporte que implica la aplicación de DWD al ámbito educacional pues permite realizar un análisis cualitativo del rendimiento académico de los estudiantes, además de considerar el ámbito contextual de cada organización educacional pues no es lo mismo clasificar a un estudiante bueno en una organización A que en una organización B. Lo anterior debido a consideraciones como el puntaje de ingreso y las calificaciones que obtiene un estudiante.

Desde este punto de vista la principal contribución del presente trabajo es mostrar un diseño de DWD simple junto a una implementación que permite apreciar las ventajas de la aplicación de un DWD para la toma de decisiones en instituciones educacionales. Además, la propuesta de DWD incluye tres elementos difusos que son difusidad entre niveles (relación difusa), difusidad en las medidas, difusidad de un nivel, y hechos difusos que se explicaran en la sección de preliminares.

El artículo está organizado de la siguiente forma: primero se presenta una sección de Preliminares que permite explicar los conceptos básicos acerca de los elementos difusos que se incorporan en la propuesta para análisis de datos de estudiantes, luego se presenta una sección con los Trabajos Relacionados y posteriormente se presenta una sección con la Metodología de Trabajo que explica el caso de estudio mediante un esquema conceptual y esquema lógico del DWD propuesto. Luego se presentan los Resultados a las consultas del DWD. Finalmente se presenta la Conclusión y Trabajos Futuros que incluyen comentarios sobre los resultados obtenidos, y posibles trabajos futuros.

A continuación, en la sección Preliminares se explican los conceptos de difusidad del DWD que se ha implementado para aplicar el análisis de datos educacionales.

PRELIMINARES

En este artículo se ha considerado la siguiente definición para la implementación del DWD *"Un DW que permite almacenar y operar medidas difusas tipo 1 y tipo 2, relaciones difusas entre niveles y niveles difusos tipo 2"*. Además, se permite agregar un grado de posibilidad a los hechos y operar con estos.

Relación difusa

Cuando una relación entre niveles es difusa, cada instancia de nivel inferior estará asociada a más de una instancia del nivel superior, generando una relación muchos a muchos, entre el nivel padre y el nivel hijo. Este tipo de jerarquía es no estricta [11]. La Figura 1 muestra un ejemplo de esquema de dimensión difusa donde se ha definido la relación entre alumno y nivel avance como difusa representada por una línea punteada. Que la relación sea difusa implica que hay asociado un grado de pertenencia entre las instancias de dimensión que participan en dicha relación. Por ejemplo, un alumno puede ser asociado a más de un nivel de avance, esto es muy frecuente debido a que los estudiantes se atrasan por reprobación de asignaturas y de esta forma quedan en distintos niveles (semestres) con asignaturas de diversos niveles sin estar asociados a un solo nivel. Entonces se debería asociar un grado de pertenencia a la relación como se muestra en la Figura 2, donde se observa que el estudiante A1 está asociado al nivel 1 (Niv1) en 0,7 grado de pertenencia y asociado al nivel 2 (Niv2) en 0,3 grado de pertenencia. Es importante indicar que el grado de pertenencia se debe obtener desde los datos y suma 1.

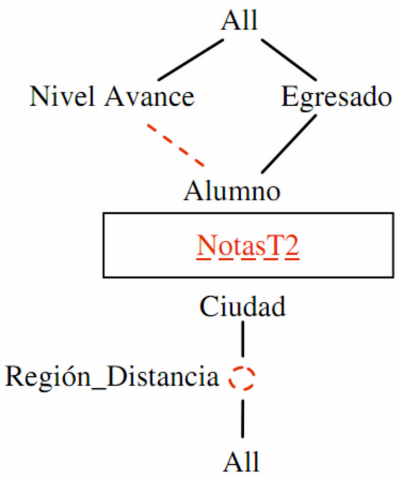

Figura 1. Esquema de DWD con relación difusa, nivel difuso tipo 2 y medida difusa tipo 2.

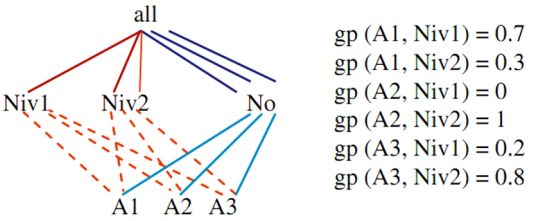

Figura 2. Extracto de instancia de dimensión para la relación difusa entre estudiante y nivel.

Medidas difusas tipo 1 y tipo 2

Cuando una medida es difusa en el diseño e implementación de esta propuesta se considera que puede ser difusa tipo 1 o difusa tipo 2. Si es medida difusa tipo 1 solo se le agrega una etiqueta a la medida. Si es medida difusa tipo 2 significa que se ha asociado un concepto difuso que tiene asociadas etiquetas lingüísticas que se representan por funciones de pertenencia. Por ejemplo, en la Figura 1 la medida notas T2 es difusa tipo 2 y su concepto difuso es calidad de las notas para lo que se definen tres etiquetas para clasificar las notas en mala, regular o buena como se muestra en la Figura 3.

Nivel difuso

Para el caso de nivel difuso tipo 2 también se agrega un concepto difuso que tiene funciones de pertenencia que contendrán las etiquetas. Por ejemplo, en el caso de Región_Distancia que se muestra en la Figura 1 el concepto difuso es la localización donde las etiquetas serían cerca y lejos (cada una con un grado de pertenencia dependiendo de la distancia a algún lugar).

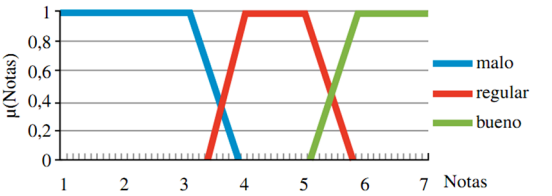

Figura 3. Funciones de pertenencia trapezoidales para el concepto calidad de las notas y las etiquetas lingüística mala, regular y buena definida sobre el atributo notas.

VENTAJAS DE UN DWD APLICADO A DATOS EDUCACIONALES

Data Warehouse difuso	Data Warehouse tradicional
Permite realizar la clasificación dependiendo del contexto, es decir, la etiquetación depende de la organización y los parámetros que ellos consideran para realizarla clasificación por medio de funciones de pertenencia.	No permite realizar ni clasificación, ni etiquetación de los datos de los estudiantes.
Permite realizar un análisis cualitativo de los datos de los estudiantes. Además del cuantitativo.	Solo permite realizar análisis cuantitativo.
Permite que una instancia pertenezca a más de una categoría debido a que se extiende la función de agregación a agregación difusa, por lo tanto, si un estudiante se atrasa en una carrera existe la opción de asociarlo a todos los niveles donde cursa asignaturas.	Solo permite asociar un estudiante a una categoría dada una instancia de categoría.

TRABAJOS RELACIONADOS

En la literatura se encuentran pocos trabajos relacionados con diseños y/o implementaciones de DWD, al contrario del caso de DW tradicional [17, 26]. Lo anterior se debe a que el proceso de implementación requiere de un esfuerzo técnico que toma un tiempo considerable y además el concepto de DWD ha sido incluido desde hace poco por el ámbito académico y no existe un estándar para su definición. En este ámbito encontramos los siguientes trabajos que aplican DWD [6, 7, 15] que se explican a continuación.

En [15] se utiliza un modelo multidimensional difuso para analizar datos financieros. En este caso se incluye la explicación del modelo por medio de su representación lógica y luego se muestran los resultados a consultas donde queda reflejada la utilidad del modelo. Este es un caso de utilización de DWD, pero no tiene relación con el ámbito educacional.

Por otro lado, en [7] se muestra la utilidad del uso de un DWD en el ámbito de análisis de indicadores y métricas web como por ejemplo páginas vistas por un usuario. Para lo cual presentan un esquema de DWD que incluye metatablas que soportan los elementos difusos en dimensiones y hechos. Además, describen cómo aplicar la operación Fuzzy Slicing y Fuzzy Dicing para obtener resultados; sin embargo, no muestran resultados gráficos a las consultas. Luego en [6] el mismo autor aplica su enfoque de DWD que utiliza dimensiones difusas y hechos difusos a datos de clientes de una fábrica de instrumentos, este trabajo está basado en la misma estructura mostrada en [7]. La formalización de la estructura de DWD que se utiliza en [6-7] se encuentra en [8] trabajo que explica cómo incorporar elementos difusos a un DW para convertirlo en DWD usando estructuras de metatablas.

En [21] se describe cómo aplicar la metodología de Kimball que se utiliza para el diseño de un DW tradicional a una extensión de dicha metodología para la construcción de un DWD. También muestran un breve caso de estudio aplicado a datos de ventas de productos y sus clientes.

Respecto a trabajos relacionados al ámbito educacional que apliquen DWD se han propuesto los siguientes trabajos [1-2, 27].

En [2] se presenta un mecanismo basado en MDA [16] que muestra cómo transformar un esquema multidimensional difuso desde el nivel conceptual al nivel lógico usando como caso de estudio un ejemplo aplicado al análisis del tiempo promedio de titulación de un estudiante, pero no indica cómo implementar consultas ya que no es el objetivo, en este trabajo solo se incorporan medidas difusas.

En [1] se presenta un trabajo que tiene como principal contribución el diseño e implementación de un modelo multidimensional difuso para administrar datos que provienen desde repositorios de objetos de aprendizaje para lo que se muestra el modelo y los esquemas estrella y copo de nieve de los casos que analizan. Sin embargo, no existe implementación ni resultado a consultas.

En [27] se presenta un trabajo cuyo objetivo es utilizar el enfoque MDA [16] para el diseño de un DWD en este trabajo se aplica el diseño a un caso de estudio de análisis de datos educacionales, pero no se muestra la implementación ya que fue declarada como trabajo futuro. En la presente investigación las

autoras muestran la implementación enfocada en poder realizar un análisis de datos de origen educacional.

Según el análisis de trabajos relacionados se observa que la implementación de un DWD para el análisis de rendimiento académico estudiantil es una contribución pues se puede tomar como modelo a aplicar en este tipo de organización y permite apoyar el proceso de toma de decisiones otorgando información cualitativa de los estudiantes.

A continuación, la sección de Metodología muestra el proceso que se llevó a cabo para la implementación del DWD, el diseño está basado en un trabajo anterior de las autoras [27].

METODOLOGÍA

Para realizar el diseño multidimensional de un DW existen enfoques impulsados por la oferta, enfoques impulsados por la demanda y enfoques híbridos [4].

El enfoque impulsado por la oferta es también conocido como enfoque impulsado por los datos. En este caso el proceso se inicia con el modelado del DW desde un análisis detallado de las fuentes de datos para determinar los elementos del DW como hechos y dimensiones posibles de considerar con los datos que se tienen. La información que se considera en los hechos representa medidas para los procesos de negocio y busca responder preguntas como: ¿Cuál es la asignatura que más reprueban los estudiantes de cierta carrera? Desde este punto de vista las dimensiones representan el marco para el análisis de estas medidas [23].

El enfoque impulsado por la demanda también es conocido como enfoque impulsado por los requisitos o dirigido por objetivos. En este caso el proceso de diseño comienza por la determinación de los requerimientos desde las necesidades de los usuarios. Luego se crea el diseño multidimesional según los objetivos seleccionados. Las ventajas de este enfoque son que pueden apoyar el proceso de reestructuración de los procesos de negocio. Este enfoque permite responder preguntas como ¿es posible cumplir con el objetivo X? [19].

El enfoque hibrido permite combinar el enfoque impulsado por los datos con el enfoque impulsado por la demanda [19].

Por otro lado, para el diseño de un DWD se han propuesto los siguientes trabajos [2, 8, 21] que ya se explicaron en la sección de Trabajos Relacionados.

Para realizar el análisis de datos educacionales mediante DWD se llevó a cabo un enfoque basado en MDA que ya fue introducido en un trabajo anterior de las autoras [27]. Según los enfoques antes explicados este paradigma es orientado a los datos; es decir, producto de un análisis exhaustivo de los datos se determinó que elementos eran posibles de implementar en el DWD, además se consultó con una experta en educación sobre la utilidad de los resultados obtenidos en el contexto educacional, esta persona también ayudó en la definición de objetivos. Sin embargo, aún queda trabajo por realizar bajo el paradigma orientado a la demanda. En el proceso se incluyeron las siguientes tareas:

- Análisis de las fuentes de datos: en esta etapa se estudiaron las fuentes de datos para determinar el diseño del DWD, es decir, qué dimensiones se podían implementar con los datos existentes, qué medidas y además que relaciones entre los datos podían ser difusas además de qué niveles y medidas podían ser difusos.
- Consulta con especialista en educación: en esta etapa se validó el análisis de las fuentes con una especialista en educación que actuó como usuario para determinar que los objetivos de análisis tenían coherencia y utilidad en el ámbito educacional.
- Diseño conceptual del DWD: en esta etapa se diseña el esquema conceptual.
- Diseño lógico del DWD: en esta etapa se diseña el esquema lógico.
- Implementación: en esta etapa se realizó el proceso ETL para cargar los datos en el cubo y en las metatablas que soportan los elementos difusos de la propuesta de DWD.

A continuación, se muestran los esquemas conceptuales y lógicos para el Cubo Indicadores de Estudiantes [27]. Posteriormente se muestran los resultados a consultas.

Cubo Indicadores de Estudiantes

El Cubo Indicadores de Estudiantes está compuesto por seis dimensiones de análisis que son ingreso, carrera, tiempo, estudiante, localización y curso. Para la implementación se utilizaron los datos de pregrado de una universidad chilena. El propósito de

implementación de estos indicadores es proporcionar información de la carga promedio de los estudiantes y sus notas promedio, haciendo un análisis por ingreso, carrera, estudiante, localización y curso. En este contexto se definieron los siguientes elementos difusos:

- Medidas difusas: que son carga promedio y notas promedio.
- Nivel difuso: que es región.
- Relación difusa: entre las categorías estudiante y nivel estudiante.

Esta última relación es difusa por esencia dado que regularmente los alumnos no se encuentran al día en el avance de su malla curricular, sino que, por el contrario, tienen asignaturas de distintos niveles (semestres) lo que nos indica que su grado de pertenencia a un nivel de su carrera es difuso.

En el caso de la dimensión localización, se observa que tiene dos niveles que son ciudad y región, donde el nivel región ha sido marcado como difuso, esto indica que se puede etiquetar como atributo difuso *tipo 2*.

En la Figura 4, se aprecia el Esquema Conceptual para el Cubo Indicadores de alumnos, se representa mediante una instancia de metamodelo CWM OLAP Difuso presentado en el trabajo de Zambrano, Varas y Urrutia [27]. En esta figura, se puede observar el estereotipo *<<FMT2>>* que representa las medidas difusas *tipo 2*, el estereotipo *<<FuzzyLevelT2>>* que representa al nivel difuso *tipo 2* y el estereotipo *<<FHLA>>* que representa a las relaciones difusas en una jerarquía.

Luego se identifican los cubos, medidas, dimensiones, atributos de dimensiones, niveles, atributos de niveles, clases de asociación y los atributos difusos *tipo 1*, *tipo 2* en medidas, niveles difusos *tipo 2* y relaciones difusas de la instancia del metamodelo CWM OLAP Difuso [14, 16, 27].

En la Figura 5 se aprecia el esquema lógico para los indicadores de alumnos carga promedio y notas

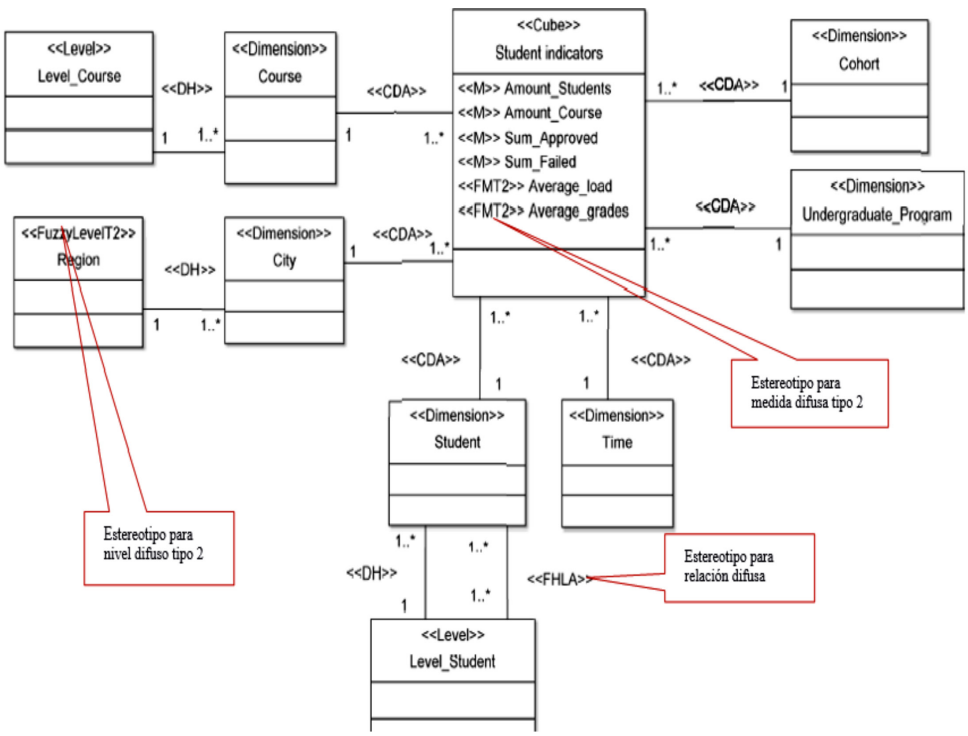

Figura 4. Esquema Conceptual para el Cubo Indicadores de alumnos con elementos difusos. estereotipados <<FMT2>> para medida difusa tipo2, <<FuzzyLevelT2>> para nivel difuso tipo2, <<FHLA>> para relación difusa.

promedio, con las estructuras para soportar los elementos difusos.

A continuación, se presentan los resultados a consultas realizadas sobre el DWD propuesto tales como:

- Análisis usando jerarquías difusas.
- Análisis usando hechos difusos.
- Análisis usando medidas difusas etiquetadas tipo 2.
- Análisis usando niveles difusos etiquetados tipo 2.

RESULTADOS

Las siguientes consultas fueron ejecutadas sobre el DWD en un SGBD [28] para responder a los indicadores buscados. Los elementos difusos se cargaron en la etapa ETL [3, 11-13] del proceso.

Análisis usando jerarquías difusas

Indicador promedio de notas por nivel

La relación jerárquica entre el nivel Student y el nivel Level_Student puede ser difusa según el caso de estudio (ver relación estereotipada <<FHLA>> en esquema conceptual de la Figura 4). De esta forma, la consulta Promedio de notas por nivel es una consulta que entrega el promedio de notas por cada nivel y además calcula el grado de posibilidad de cada hecho resultante de la agregación debido a que cada estudiante tiene una relación difusa con los niveles.

Por un lado, en la Tabla 1 se presenta los resultados del indicador "Promedio de notas por nivel", se puede apreciar que solo se entrega como resultado los niveles 101 al 302 (los primeros tres años), esto debido a que no hay notas que tengan un grado de relación entre Student y Level_Student que supere el umbral $\alpha = 0{,}3$ utilizado en la consulta.

Por otro lado, en la Tabla 2 se muestran los resultados de la misma consulta, pero utilizando la jerarquía no difusa entre Student y Level_Student, donde el nivel del estudiante es considerado como el nivel de la asignatura menos avanzada que cursa. Como se puede apreciar los resultados son distintos, en el caso de la consulta difusa los promedios son generalmente más bajos, esto puede ser debido a que en tal caso se consideran todos los niveles

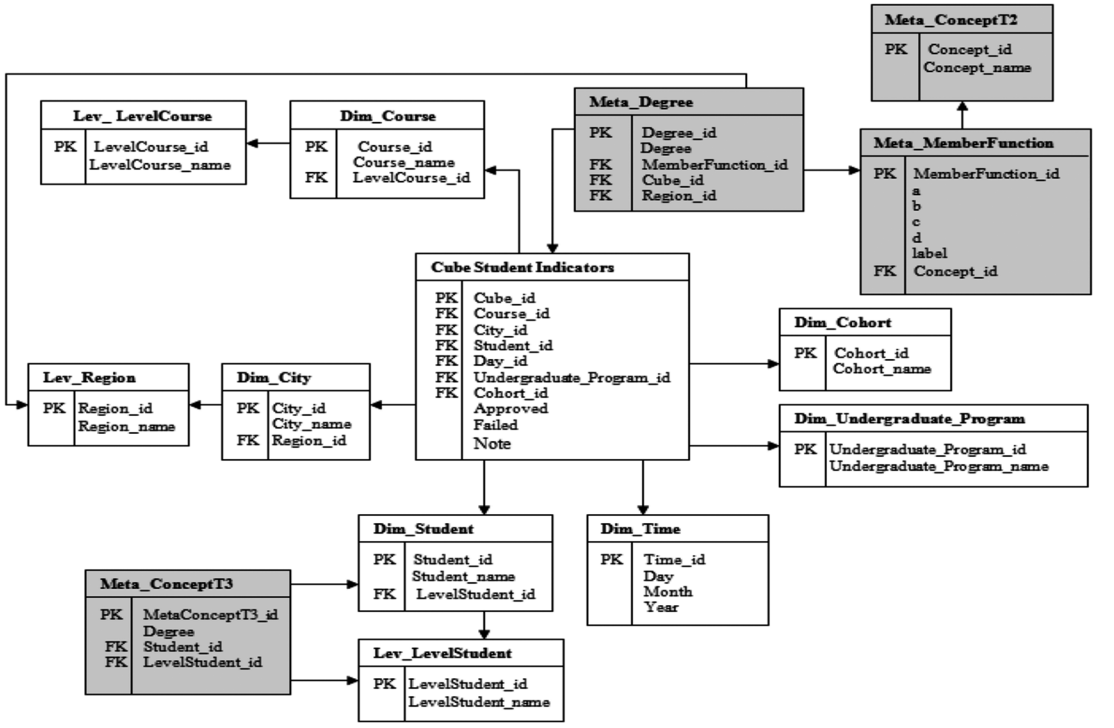

Figura 5. Esquema lógico para los indicadores de alumnos carga promedio y notas promedio.

Tabla 1. Resultado de la consulta para indicador "Promedio de Notas por Nivel", utilizando los parámetros $\alpha = 0{,}3$, $\beta = 1$, $\delta = 1$ y la relación difusa entre Student y Level_Student.

Student_Level	Avg_Nota	δ
101	3,71	0,30
102	4,06	0,30
201	3,96	0,30
202	3,82	0,32
301	4,29	0,30
302	4,53	0,32

Tabla 2. Resultado de la consulta "Promedio de notas por nivel", utilizando la relación NO difusa entre Student y Level_Student.

Student_Level	Avg_Nota
101	3,15
102	4,36
201	4,50
202	4,07
301	4,31
302	4,59
401	5,27
402	5,09
501	4,98
502	5,24
601	4,62
602	5,33

del alumno que tengan una relación superior a 0,3, considerando, por lo tanto, también las notas de los primeros niveles los que contienen más reprobaciones. Además, se puede apreciar que la posibilidad de los hechos es entorno al 0,3 para los hechos, esto se puede interpretar como que, para cada nivel, los hechos obtenidos producto de la agregación de Avg_nota son muy dispersos, teniendo en su valor mínimo un 0,3 de posibilidad de ser el valor indicado.

Análisis sobre hechos difusos

Promedio de notas y coeficiente de aprobación por carrera

El modelo propuesto tiene la capacidad de asignar un nivel de posibilidad al conjunto de hechos. Este parámetro está asociado independientemente a cada hecho y puede ser usado como criterio al momento de seleccionar los hechos que deben ser agregados.

El siguiente ejemplo realiza una consulta en el cubo con tal de obtener el promedio de notas de los alumnos por cada carrera, que pertenezcan al ingreso 2010 para la asignatura CÁLCULO I.

La Tabla 3 muestra los resultados de la aplicación de una consulta de agregación no difusa estableciendo el parámetro $\delta = 1$ con tal de considerar solo hechos totalmente posibles.

Tabla 3. Resultado de la consulta "Promedio de notas y coeficiente de aprobación por carrera", utilizando $\alpha = 1$, $\beta = 1$, $\delta = 1$.

Undergraduate Program	Avg_ Nota	Ratio_ Reprob	Ratio_ Aprob	$\delta = 1$
Geología	4,20	0,28	0,72	1
Ingeniería Civil en Computación e Informática	3,25	0,48	0,52	1
Ingeniería Civil en Metalurgia	3,41	0,46	0,54	1
Ingeniería Civil en Minas	3,78	0,34	0,66	1
Ingeniería Civil Industrial	3,43	0,44	0,56	1

Al realizar la consulta con el parámetro $\delta = 0{,}9$ se puede observar en la Tabla 4 un leve cambio en el promedio y coeficientes de aprobación y reprobación, cambiando además la posibilidad del hecho agregado. Por otro lado, si el parámetro se establece $\delta = 0{,}5$ el resultado vuelve a cambiar como se aprecia en la Tabla 5.

Es importante indicar que si se prueba con más fuentes se podrá establecer precisiones o posibilidades en los hechos de forma más real.

También se puede señalar que el modelo sería muy útil en un caso de integración de fuentes de datos externas donde se puede asignar un grado de posibilidad a los hechos según la confianza que se tenga en la fuente de datos.

Tabla 4. Resultado de la consulta "Promedio de notas y coeficiente de aprobación por carrera", utilizando $\alpha = 1, \beta = 1, \delta \geq 0{,}9$.

Undergraduate Program	Avg_ Nota	Ratio_ Reprob	Ratio_ Aprob	$\delta \geq 0{,}9$
Geología	4,20	0,28	0,72	1
Ingeniería Civil en Computación e Informática	3,25	0,48	0,52	1
Ingeniería Civil en Metalurgia	3,41	0,46	0,54	1
Ingeniería Civil en Minas	3,78	0,34	0,66	1
Ingeniería Civil Industrial	3,37	0,44	0,56	1

Tabla 5. Resultado de la consulta "Promedio de notas y coeficiente de aprobación por carrera", utilizando $\alpha = 1, \beta = 1, \delta \geq 0{,}5$.

Undergraduate Program	Avg_ Nota	Ratio_ Reprob	Ratio_ Aprob	$\delta \geq 0{,}5$
Geología	4,20	0,28	0,72	1
Ingeniería Civil en Computación e Informática	3,25	0,48	0,52	1
Ingeniería Civil en Metalurgia	3,41	0,46	0,54	1
Ingeniería Civil en Minas	3,78	0,34	0,66	1
Ingenieria Civil Industrial	3,25	0,44	0,56	1

Análisis sobre medidas difusas etiquetadas tipo 2

La propuesta otorga la capacidad de etiquetar una medida de los hechos mediante etiquetas lingüísticas agrupadas en conceptos. Para el ejemplo tratado sobre indicadores de estudiantes se ha definido el concepto de "rendimiento" que agrupa las siguientes etiquetas lingüísticas mala, regular y buena, cada una con su función de pertenencia trapezoidal para las medidas basadas en la nota con escala del 1 al 7. La Figura 3, antes mostrada, indica las funciones definidas para cada etiqueta.

Los atributos difusos tipo 2 en el modelo propuesto pueden estar asociados a variables que componen medidas en el conjunto de hechos. De esta forma es posible realizar operaciones de selección difusa sobre estos atributos con tal de seleccionar valores de variables en los hechos de acuerdo a los criterios definidos en las etiquetas lingüísticas. Para realizar esta operación se utiliza los operadores de selección difusa definidos en [20].

A continuación, se muestran dos ejemplos de aplicación sobre medidas difusas etiquetada tipo 2.

Cantidad de notas altas por ingreso

La Tabla 6, presenta los resultados de la consulta "Cantidad de notas altas por ingreso", donde se puede apreciar que el ingreso con más cantidad de notas absolutamente buenas es el ingreso 2007 con 416 notas con posibilidad 1 de ser buenas. También se puede apreciar que otras cantidades también son posibles en menor medida, por ejemplo, para el ingreso 2013 solo hay 22 notas que son buenas con una posibilidad 0,5.

Cantidad de notas malas y regulares por carrera

En el siguiente ejemplo de la Tabla 7, se realiza una consulta similar a la consulta anterior, pero se mezclan dos etiquetas lingüísticas: mala y regular.

La consulta para obtener la cantidad de notas que son malas y regulares (ambas) se realiza por medio de la operación OWA MAM [24] entre las funciones de pertenencia para ambas etiquetas.

Como resultado, la cantidad de notas es pequeña debido a que son pocas las notas por cada carrera

Tabla 6. Resultado de la consulta "Cantidad de notas altas por ingreso".

	μ *vs count ()*	2007	2008	2009	2010	2011	2012	2013
$\mu_{buena}(nota) \otimes \mu_{buena}(nota)$	0,50	74	73	66	63	62	46	22
	0,62	84	80	67	52	63	39	25
	0,75	66	67	52	43	39	32	20
	0,87	64	48	48	30	26	39	24
	1,00	416	362	300	249	183	201	103

Tabla 7. Resultado de la consulta "Cantidad de notas malas y regulares por carrera".

	μ vs count ()	Geología	Informática	Metalurgia	Minas	Industrial
$\mu_{buena}(nota) \otimes \mu_{buena}(nota)$	0,13	7	17	5	8	19
	0,17	19	33	10	39	62
	0,25	11	25	5	22	31
	0,33	13	21	20	18	54

que son malas y regulares, dado que deben cumplir con las dos descripciones al mismo tiempo.

Análisis sobre niveles difusos etiquetados tipo 2
La propuesta soporta niveles difusos mediante la etiquetación de estos. En el ejemplo de la Figura 6 el nivel región posee un atributo distancia que es asociado al concepto distancia que posee las etiquetas lingüísticas cerca y lejos con sus respectivas funciones de pertenencia trapezoidal como se muestra a continuación.

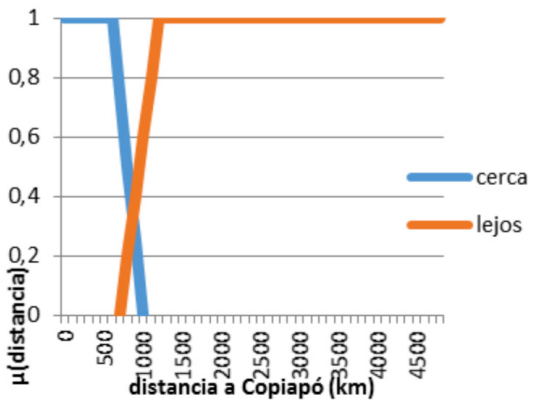

Figura 6. Funciones de pertenencia trapezoidales para el concepto distancia y las etiquetas lingüísticas cerca y lejos definidos sobre el atributo distancia del nivel región.

La etiquetación de niveles permite realizar la selección de instancias de nivel no solo mediante sus valores exactos o atributos, sino que también a partir de sus etiquetas lingüísticas.

La diferencia entre los niveles difuso tipo 2 y los hechos difuso tipo 2 es que la "difusidad" solo se puede aplicar al atributo de nivel, por lo que solo cuando se haga referencia al nivel en una consulta se podrá acceder a sus etiquetas.

Al contrario, los atributos difuso tipo 2 se mantienen independiente de la operación que realice en el cubo, siempre y cuando se mantenga la medida que tiene asociado el concepto. Es por ello, que los niveles difusos permiten un nivel de detalle y análisis más completo a la hora de realizar las operaciones dice, pero no afecta la operación roll-up o drill-down.

Cantidad de alumnos en regiones cercanas y lejanas
Cuando los niveles son etiquetados como atributos difusos tipo 2, permiten las mismas operaciones que se realizan sobre los hechos por medio de operaciones OWA, en nuestro caso específico operadores MAM y MOM [5, 24].

La Tabla 8 contiene los resultados de realizar una consulta que permite calcular la cantidad de alumnos por región, considerando solo las regiones que están lejos y cerca al mismo tiempo, es decir, las que no están definidas como totalmente posible lejos o totalmente posible cerca. De esta forma la consulta trata sobre la aplicación de la operación MAM sobre las dos funciones de pertenencia descritas anteriormente. Como se puede observar en los resultados, solo las instancias de Tarapacá a Libertador Bernardo O'Higgings se consideran en esta categoría exceptuando las que están totalmente cerca como por ejemplo la misma Región de Atacama. También se puede apreciar que la pertenencia es bastante baja en todos los casos, esto es debido a las funciones definidas, que no definen quizás correctamente este concepto, lo que se ha realizado en forma consciente para demostrar la importancia que puede tener una buena definición de funciones de pertenencia. Respecto a la definición de los parámetros de las funciones de pertenencia se puede utilizar una definición manual utilizando el criterio de un usuario que tenga relación con el contexto de análisis o se puede utilizar una propuesta para la parametrización de funciones de pertenencia para DWD mediante técnicas de multi-level thresholding como se explica en el trabajo de

Rojas, Zambrano, Varas y Urrutia [18] y de esta forma obtener los parámetros de la función de pertenencia de forma automática desde los datos. Lo anterior es muy útil debido a que permitirá obtener parámetros de la función de pertenencia de acuerdo con el contexto que está validado por los datos que se tienen de la organización, ya que no es lo mismo etiquetar a un estudiante como bueno en una universidad A, que en una universidad B. Pues, lo anterior depende de diversos aspectos, como puntaje de ingreso, motivación por la carrera, carencia de estrategias de aprendizaje, etcétera.

Tabla 8. Resultado de la consulta "Cantidad de alumnos en regiones cercanas y lejanas".

Región	$\mu_{cerca} \otimes \mu_{lejos}$	Cantidad
Región de Tarapacá	0,04	9
Región Metropolitana	0,12	62
Región de Valparaíso	0,01	31
Región del Libertador General Bernardo O'Higgins	0,28	28

Si se realiza la misma consulta, pero utilizando las funciones de pertenencia mostradas en la Figura 7, los resultados son los presentados en la Tabla 9.

En la Tabla 9 se puede también apreciar la distancia asociada a cada región, donde podemos ver que, con la nueva definición de funciones de pertenencia, se agrega la Región del Maule y de Antofagasta, como también los grados de pertenencia son ahora más altos, teniendo esto más sentido, ya que por ejemplo Antofagasta es cercano a Copiapó, pero a la vez también está lejos, mucho más que La Serena,

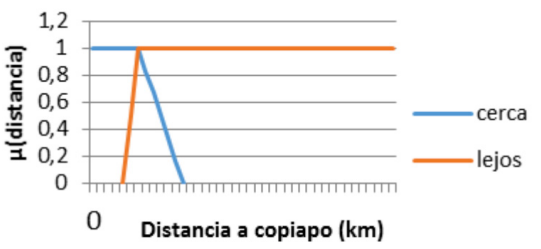

Figura 7. Definición alternativa de funciones de pertenencia trapezoidales para el concepto distancia y las etiquetas lingüísticas cerca y lejos definidos sobre el atributo distancia del nivel región.

quien cae en la categoría de cercano. Con lo anterior se muestra la importancia de una buena definición de los parámetros de la función de pertenencia, el ideal es determinarlos desde los datos [18].

Tabla 9. Resultado de la consulta "Cantidad de alumnos en regiones cercanas y lejanas" utilizando funciones de pertenencia de la Figura 5.

Región	Distancia	$\mu_{cerca} \otimes \mu_{lejos}$	Cantidad
Región de Antofagasta	574	0,87	31
Región de Tarapacá	986	0,36	9
Región Metropolitana	803	0,66	62
Región de Valparaiso	754	0,74	31
Región del Maule	1060	0,23	8
Región del Libertador General Bernardo O'Higgins	887	0,52	28

A modo de resumen es importante indicar que para llegar a la implementación presentada en este trabajo se realizaron varias etapas previas de análisis de los datos educacionales, luego se consultó con experta en educación para validar el objetivo de análisis y posteriormente en las etapas de modelamiento se diseñaron los esquemas conceptuales y esquema lógico que se usaron para esta implementación. De esta forma el trabajo siguió un procedimiento estricto que se refleja en la metodología de trabajo que está alineada con lo que sugiere [23].

Desde un punto de vista educacional la propuesta representa un caso de estudio complejo que arroja conclusiones útiles en este ámbito y cuya aplicación permite solucionar problemas que tienen un carácter relevante. La propuesta representa un modelo útil para realizar minería de datos educacional.

A continuación, se presentan las conclusiones del trabajo realizado.

CONCLUSIONES

Se ha presentado una propuesta de implementación de DWD para el análisis de datos educacionales.

El trabajo formulado incluye ejemplos de análisis para todos los casos de elementos difusos que se han propuesto en este caso de estudio, como son

medidas difusas, relaciones difusas entre niveles, niveles etiquetados tipo 2 y hechos difusos.

Es importante indicar que la experimentación permitió mostrar que la definición adecuada de los parámetros de la función de pertenencia trae consigo buenos resultados a las consultas.

Además, estos parámetros dependen del contexto, es decir, no existe una parametrización universal, sino que depende del contexto de la organización donde se aplique el análisis. Por lo anterior, lo ideal es obtener los parámetros de la función de pertenencia teniendo en cuenta los datos que se dispone como ya se planteó en el trabajo de Rojas, Zambrano, Varas y Urrutia [18].

La propuesta representa un modelo útil para realizar minería de datos sobre datos educacionales. También se puede indicar que en el proceso de toma de decisiones es natural que se piense en aspectos de análisis cualitativo por lo que la propuesta muestra por medio de los resultados del DWD un aporte en la toma de decisiones a nivel de datos educacionales en una organización.

Como trabajo futuro se trabaja en la definición de un instrumento que permita validar las necesidades de información difusa en el análisis de los datos educacionales para de esta forma complementar una metodología de análisis de datos educacionales usando DWD.

Desde un punto de vista de implementación se trabaja en generar nuevas medidas como el tiempo en que un estudiante tarda en avanzar de nivel en su carrera, etcétera.

AGRADECIMIENTOS

Se agradece al investigador Darío Rojas Díaz por su valioso aporte en esta investigación.

REFERENCIAS

[1] G. Appelgren Lara, M. Delgado and N. Marin. "Fuzzy Multidimensional Model ling for Flexible Querying of Learning Object Repositories". LNAI 8132, pp. 112-123. 2013.

[2] S. Carrera, M. Varas y A. Urrutia, "Transformación de esquemas multidimensionales difusos desde el nivel conceptual al nivel lógico". Ingeniare. Revista Chilena de Ingeniería. Vol. 18 Nº 2, pp. 165-175. Mayo - Agosto 2010.

[3] S. Chaudhuri and U. Dayal. "An overview of Data Warehousing and OLAP Technology". SIGMOD Record. Vol. 26, pp. 65-74. 1997. Pearson. 2004. ISBN: 8420540250.

[4] A. Cravero and S. Sepúlveda. "A chronological study of paradigms for datawarehouse design". Ingeniería e Investigación. Vol. 32 Nº 2, pp. 58-62, 2012.

[5] M. Delgado, C. Molina, D. Sánchez, A. Vila and L. Rodríguez-Ariza. "A fuzzy multidimensional model for supporting imprecision in OLAP". Proceedings of IEEE International Conference on Fuzzy Systems. ISBN 0-7803-8353-2. 2004.

[6] D. Fasel. "A fuzzy data warehouse approach for the customer performance measurement for a hearing instrument manufacturing company". Procceding of Fuzzy Systems and Knowledge Discovery, IEEE Explore. 2009.

[7] D. Fasel and D. Zumstein. "A Fuzzy Data Warehouse Approach for Web Analytics Visioning and Engineering the Knowledge Society". A Web Science Perspective Lecture Notes in Computer Science. Vol. 5736, pp. 276-285. 2009.

[8] D. Fasel and K. Shahzad. "A Data Warehouse model for integrating fuzzy concepts in meta table structures". In Proc. ECBS 2010, pp. 100-109. 2010.

[9] J. Galindo, A. Urrutia and M. Piatinni. "Fuzzy databases: Modeling, design and implementation". Idea Group, Inc. 2009.

[10] J. Galindo, A. Urrutia and M. Piattini. "Fuzzy Databases: Modeling, Desing and Implementation". Idea Gruop Publishing Hershey, USA. 2004.

[11] C. Hurtado and C. Gutiérrez. "Data Warehouses and OLAP: Concepts, Architectures and Solutions". Chapter Handling Structural Heterogeneity in OLAP. Idea Group, Inc. 2007.

[12] W. Inmon. "Building the Data Warehouse". John Wiley & Sons. 2002.

[13] R. Kimball, M. Ross and R. Merz. "The Data Warehouse toolkit: The complete guide to dimensional modeling". John Wiley & Sons. 2002.

[14] J.-N. Mazón and J. Trujillo. "An MDA approach for the development of data warehouses". Decision Support Systems Vol. 45, pp. 41-58. 2008.

[15] C. Molina, M. E. Gómez, J. M. Torre and M.A. Vila Miranda. "Using Fuzzy DataCube for Exploratory Analysis in Financial Economy". EUSFLAT Conf., pp. 424-429. 2005.

[16] OMG. "MDA guide". Versión 1.0.1. URL: http://www.omg.org/docs/omg/03-06-01.pdf. June, 2003.

[17] G. Poblete y C. Zambrano. "Bases de Datos multidimensionales para datos educacionales". Jornadas Chilenas de Computación. Temuco, Chile. 2013. URL: http://jcc2013.inf.uct.cl/wp-content/ proceedings/ECC/Bases%20de%20Datos%20multidimensionales%20para%20datos%20educacionales.pdf

[18] D. Rojas, C. Zambrano, M. Varas and A. Urrutia. "A Multi-Level Thresholding-Based Method to Learn Fuzzy Membership Functions from Data Warehouse". CIARP, pp. 664-674. 2011.

[19] O. Romero and A. Abello. "A survey of Multidimensional Modeling Methodologies". International Journal of Data Warehousing & Mining. Vol. 5 Nº 2, pp. 1-23. 2009.

[20] E. Rundensteiner and L. Bic. "Aggregates in posibilistic databases". In Proceeding of the 15th ConJ in Very Lnrge Databases (VLDB'89). Amsterdam, Holland, pp. 287-295. 1989.

[21] L. Sapir, A. Shmilovici and L. Rokach. "A methodology for the design of a fuzzy Data Warehouse". 4th IEEE Conference on Intelligent Systems. Bulgaria. 2008.

[22] A. Urrutia. "Definición de un modelo conceptual para bases de datos difusas". Tesis para optar al grado de doctor. Universidad de Castilla-La Mancha. España. 2003.

[23] R. Winter and B. Strauch. "A method for demand driven information requirements analysis in data warehousing projects". Proceedings of the 36th Annual Hawaii International Conference on System Sciences, pp. 9-14. 2003.

[24] R. Yager. "Aggregation Operators and fuzzy systems modeling". Fuzzy sets and Systems 67, pp. 129-145. 1994.

[25] L.A Zadeh. "Fuzzy Sets". Information and Control. Vol. 8, pp. 338-353. 1965.

[26] C. Zambrano, D. Rojas, K. Carvajal and G. Acuña. "Análisis de rendimiento académico estudiantil usando Data Warehouse y Redes Neuronales". Ingeniare. Revista Chilena de Ingeniería. Vol. 19 Nº 3, pp. 369-381. 2011.

[27] C. Zambrano, M. Varas y A. Urrutia, "Enfoque MDA para el diseño de un data warehouse difuso". Ingeniare. Revista Chilena de Ingeniería. Vol. 20 Nº 1, pp. 99-113. Abril 2012.

[28] A. Urrutia "Implementación de Businees Intelillence en plataforma Free de Pentaho". Aplicaciones Posgress, Wekw y Kettle. Editorial Académica Española. ISBN: 978-3659-06423-4. Madrid, España. Mayo 2013.

[29] J. Galindo, A. Urrrutia and M. Piattini. "Fuzzy Databases: Modeling, Desing and Implementation". Editorial Idea Gruop Publishing Hershey (IGI Publishing) USA. DOI: 10.4018/978-1-59140-324-1, ISBN13: 9781591403241, 2006.

[30] A. Urrutia and J. Galindo. "Fuzzy Database Modeling: An Overview and New Definitions. Anbumani". K. (eds.), Soft Computing Applications for Database Technologies. Information Science. Soft Computing Applications for Database Technologies: Techniques and Issues edited by Drs. K. Anbumani and R. Nedunchezhian. Editorial IGI Global Publication. USA. 2010.