resumen

Se presentan una estrategia y un framework asociado que usan minería de datos para identificar los metadatos extensionales requeridos en la optimización de consultas en Organizaciones Virtuales (OV).

Generación de metadatos extensionales en organizaciones virtuales

Diego Ardila Álvarez
Natalia Valencia Lesmes
José Abásolo Prieto
María del Pilar Villamil

introducción

Las Organizaciones Virtuales (OV) son un modelo de organización en las que entidades autónomas se reúnen para colaborar y compartir recursos [11]. Las OV pueden ser clasificadas de diferentes formas; por ejemplo, de acuerdo con el número de fuentes y los recursos que comparten. Este trabajo se orienta a las OV que funcionan bajo un esquema federado y comparten información a gran escala. Es decir, aquellas que se caracterizan por ser sistemas altamente distribuidos, heterogéneos y encargados de manejar un gran número de fuentes de un mismo dominio con grandes volúmenes de datos.

En este tipo de OV la información es el recurso compartido más importante, por lo tanto es indispen-

sable disponer de mecanismos para su búsqueda y acceso. Para ello es usual contar con un mediador responsable de realizar el procesamiento de consultas. Uno de los problemas que afronta este mediador es identificar las fuentes que tienen información relevante para responder a una consulta y así no incurrir en los altos costos que significaría buscar en todas las fuentes.

Diferentes aproximaciones se han propuesto para resolver esta problemática. La mayoría de los trabajos se apoyan en la existencia de **Metadatos Intencionales**. Esta clase de metadatos contiene la descripción sobre los conceptos manejados por cada fuente. Por ejemplo, en el caso de una fuente relacional, estos metadatos corresponden al esquema de la base de datos. Aunque este mecanismo permite filtrar fuentes, en OV donde las entidades pertenecen a un mismo dominio, el número de entidades seleccionadas podría ser grande comparado con las fuentes que efectivamente poseen información para resolver la consulta. Para mejorar la selección de fuentes candidatas, en [14] se propone el uso de metadatos extensionales. Los **Metadatos Extensionales** corresponden a información asociada a las instancias manejadas por una fuente. Estos pueden ser obtenidos mediante por lo menos tres formas: 1) A través de su declaración explícita por parte de algún responsable de la fuente; 2) Aprovechando los resultados de consultas previas; y, 3) Utilizando técnicas de minerías de datos. En este artículo se explora la última forma.

La utilización de minería de datos presenta retos en un contexto de gran escala, puesto que las estrategias de aplicación tradicionales requieren de la participación de especialistas del dominio y de expertos en procesos de extracción de conocimiento específicos a cada fuente. Con el fin de superar estos retos, este trabajo presenta un *framework* escalable que utiliza minería de datos para identificar los metadatos extensionales que apoyan la selección de fuentes en OV. Adicionalmente, describe una estrategia para la construcción de dicho framework.

La estructura del artículo se describe a continuación. La sección 2 introduce la estrategia propuesta para hacer escalable un proceso de minería de datos tradicional. La sección 3 presenta los resultados de aplicar la estrategia en el segmento de entidades hospitalarias. La sección 4 describe un framework para la generación de los metadatos extensionales que se basa en dicha estrategia. La sección 5 ex-

pone un prototipo funcional del framework descrito. Y, finalmente, la sección 6 presenta las conclusiones y perspectivas de investigación.

1. Aproximación a la problemática: estrategia propuesta

Para brindar una solución escalable al problema de la caracterización de fuentes de datos en una OV a gran escala, se propuso una aproximación inductiva. En esta aproximación se identifican tres pasos fundamentales: 1) la segmentación de las fuentes; 2) la generación de metadatos extensionales en fuentes representativas de cada segmento; 3) la caracterización y generalización del proceso de generación de metadatos, de modo que este pueda ser replicado a otras fuentes de los segmentos analizados con poca o nula intervención humana.

El objetivo de la segmentación inicial es identificar grupos de fuentes que pueden ser diferenciadas, mediante un conjunto de características comunes. La segmentación se realiza de acuerdo con el criterio dado por expertos del dominio y con los requerimientos de consulta de la OV. Por ejemplo, en una OV del sistema colombiano de seguridad social, el resultado de una segmentación podría establecer como segmentos las Entidades Gubernamentales, las EPS y las IPS. En segundo lugar, se seleccionan una o más fuentes representativas del segmento del que se quiere generar metadatos extensionales. Sobre dichas fuentes se aplica un proceso de minería de datos descriptiva para establecer la forma de los metadatos y caracterizar el proceso utilizado para ello; una posible forma de los metadatos es, retomando el ejemplo anterior, el tipo de servicios ofrecido por una IPS.

Finalmente, se realiza la conceptualización con el fin de determinar cómo puede ser replicada la generación de metadatos de forma automática en otras fuentes. Esto equivale a formalizar de manera explícita el conocimiento utilizado en el proceso tradicional a través de bases de conocimiento u ontologías [7], tal como se propone en [12] y [13]. En este orden de ideas, la conceptualización es utilizada como un insumo para la construcción de un sistema que automatice el proceso. Los detalles de la misma dependen de la metodología seguida durante el proceso de minería. En particular, para uno que utilice CRISPDM [4], esta

puede consistir en formalizar el conocimiento utilizado en cada una de las etapas. Ejemplos de conocimiento conceptualizable para la metodología mencionada se muestran en la Tabla 1.

Fase	conocimiento conceptualizable.
Entendimiento del Negocio	Conceptualización del dominio del segmento. Es posible reutilizar ontologías existentes.
Entendimiento de los Datos	Identificación de los datos necesarios para la generación de metadatos. Formalización de los datos y reglas de negocio que permitan establecer atributos reduldantes y/o faltantes.
Preparación de los Datos	Conceptualización de atributos derivados útiles para el proceso.
Modelamiento	Diseño de sesiones de minería: conjunto de algotitmos utilizados y parametrizaciones requeridas.
Evaluación	Reglas para descartar modelos que no tengan sentido en el dominio.
Despliegue	Reglas de transformación de modelos a metadatos que sirvan para identificar fuentes con información relevante en el procesamiento de consultas.

Tabla 1: Ejemplos de conomiento conceptualizable en metodología crispdm

Esta estrategia fue aplicada en un caso de estudio del sector salud que sirvió para proponer un framework de automatización. A continuación se describe el caso de estudio y en la sección 4 se presentan las características del framework.

2. Definición de Metadatos extensionales en una fuente particular: caso de estudio

El caso de estudio aplicó la estrategia propuesta, exponiendo resultados para el segmento de las Instituciones Prestadoras de Salud (IPS).

Los metadatos generados consistieron en los perfiles de las hospitalizaciones que puede atender una fuente en este segmento. Cada perfil es un clúster que, además de su tamaño e identificador, presenta los sistemas que estuvieron comprometidos durante la hospitalización (SC), la causa de la hospitalización (CH), la función que cumplieron los principales medicamentos (FM) en la misma y la finalidad de los procedimientos practicados (FP).

En este sentido, la forma de un metadato del segmento de las IPS corresponde a un predicado de la forma: Tamaño=X ^ RangoEdad=Y ^ SCPrimario=Z ^ SCSecundario=W ^ FP=U ^ CH=V ^ FM=Q. En la tabla 2 se muestran algunos

de los valores obtenidos para estas variables (X, Y, Z, W, U, V y Q) durante el caso de estudio.

Posibles Perfiles IPS	1	2	3
Tamaño	50%	17%	33%
Rango de Edad	Adulto	Niño	Adulto
Sexo	F	M	M
SC Primario	Cardíaco	Cardíaco	Cardíaco
SC Secundario	Ninguno	Cardíaco	Cardíaco
FP	Diagnóstico y Terapéutico	Diagnóstico y Terapéutico	Diagnóstico y Terapéutico
CH	Enfermedad General	Enfermedad General	Enfermedad General
FM	No Reportada	No Reportada	No Reportada

Tabla 2: Ejemplo metadatos extensionales: perfiles de una ips

Toda vez que los perfiles contienen información sobre el tipo de hospitalizaciones que atiende la fuente, estos pueden ser usados por un mediador para estimar si la fuente tiene o no información relevante para responder a una consulta afín a los metadatos, tal como se describe en [1].

Para obtener este tipo de metadatos, se siguió un proceso de minería de datos tradicional en una fuente del segmento utilizando la metodología CRISPDM [4]. Como etapas críticas del proceso y por ende, susceptibles a conceptualizar, se tuvo el perfilamiento de los datos, la selección y ejecución de algoritmos y la construcción del conjunto de datos[1] [2]. A continuación se describen sus principales características.

Perfilamiento de los datos

Las fuentes de datos utilizadas fueron los archivos RIPS [3] que reportan eventos de tipo hospitalización del período comprendido entre los años 2003 y 2006. A partir de cada registro de dichos archivos se obtuvieron los siguientes campos de una hospitalización: un diagnóstico principal, un diagnóstico secundario y un conjunto de procedimientos y medicamentos suministrados. La exploración de los datos en la etapa de perfilamiento permitió encontrar una gran cantidad de nombres de medicamentos no válidos y de hospitalizaciones sin medicamentos reportados; en este caso se decidió prescindir de los respectivos registros para el proceso de minería.

1 Model set en inglés

Construcción del conjunto de datos

En la construcción del conjunto de datos se trabajó con especialistas que ayudaron a identificar atributos derivados, abstrayendo taxonómicamente los campos mencionados. Para ello se aprovecharon las taxonomías construidas en [6]. A cada diagnóstico se le asignó uno de los sistemas fisiológicos afectados por la enfermedad encontrada en el paciente; por ejemplo, la cirrosis hepática fue mapeada al sistema hepático. Del mismo modo, para cada medicamento y cada procedimiento, se derivó su función y finalidad respectivamente. Así pues, una instancia del registro construido tuvo como columnas cada uno de los elementos de alto nivel de la taxonomía. En este sentido, las taxonomías sirvieron para dos propósitos: el primero, obtener segmentos que representaran el perfil global de la fuente sin incurrir en las particularidades de cada evento; el segundo, mitigar los efectos de la maldición de la dimensionalidad [8].

Selección y ejecución de algoritmos

Se utilizaron los algoritmos K-Means (KM) y Farthest First (FF) [10]. Los clusters de KM permitieron encontrar los casos comunes en la fuente, mientras que los de FF los casos especializados.

3. Propuesta de framework para generación de metadatos extensionales

El caso de estudio fue la base para proponer OBME – Ontology Based Metadata Extractor -, un framework escalable y semi-automático para la generación de metadatos extensionales.

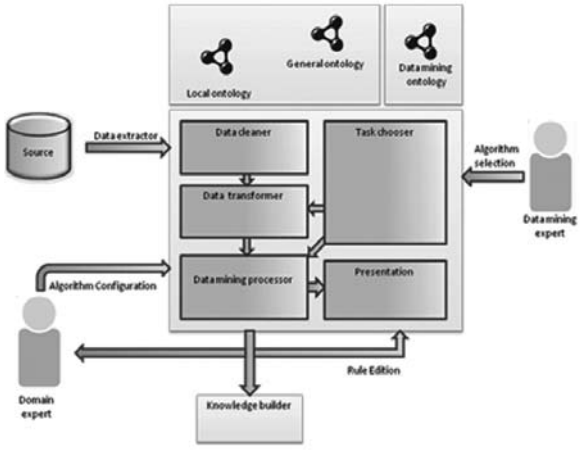

Figura 1. Arquitectura OBME

OBME está compuesto por dos tipos de componentes que son ilustrados en la figura 2: los componentes funcionales y el conjunto de bases de conocimiento. El primer tipo de componentes corresponde a los módulos que implementan las diferentes etapas del proceso de generación de metadatos extensionales y que fueron identificados durante el proceso de generación de metadatos de la IPS. Su descripción se realiza a continuación:

El **task chooser** es responsable de determinar los algoritmos que deben ser ejecutados para generar los metadatos. Entre los factores que afectan esta tarea se tiene el dominio específico de la fuente, el área de la cual tiene información y los tipos de usuarios que podrían acceder a ésta; como en el caso de estudio, los algoritmos podrían ser KM y FF, sin embargo, dependiendo de los factores mencionados, podrían haber otros más apropiados. El **data extractor** consulta la información mínima necesaria para generar los metadatos extensionales; para este propósito, obtiene la información de la fuente y la convierte en términos de un lenguaje canónico que pueda ser manipulado por el sistema. El **data cleaner** realiza un proceso de perfilamiento donde se trata de determinar la calidad de los datos a partir de reglas de negocio globalmente conceptualizadas; la finalidad de este módulo es obtener metadatos adicionales que permitan a la OV tener entre sus parámetros de selección la calidad de los datos que se almacenan. El **data transformer** es responsable de realizar el pre-procesamiento de los datos y construir el conjunto de datos con base en las decisiones tomadas por el **task chooser**; una posible tarea que este módulo incluye es la obtención de atributos derivados. Finalmente, el módulo **data mining processor** se encarga de ejecutar los algoritmos seleccionados.

El segundo tipo de componentes corresponde al conjunto de ontologías que soportan la ejecución de los módulos recién descritos. Estas, además de contener información que describe la estructura de la fuente (metadatos intencionales), reúnen parte del conocimiento que expertos en minería de datos y expertos del dominio suministrarían para el análisis de la misma. La idea es conceptualizar el conocimiento necesario acerca del proceso de generación de metadatos, de modo que pueda ser ejecutado sin la intervención de los expertos, abordando así el problema de escalabilidad. Para ello, es necesario realizar casos de estudio en el segmento de fuentes de interés, siguiendo la estrategia descrita

anteriormente. El desarrollo de estas bases de conocimiento se encuentra en investigación y hace parte del trabajo futuro de este proyecto.

Finalmente, es de notar que si bien el objetivo es reducir la intervención humana, también se incluye un módulo de presentación de resultados. De esta forma, el sistema funciona en modo automático y en modo experto para dar transparencia al proceso. Esto es fundamental puesto que el sistema no deja de ser parte de un proceso de descubrimiento de conocimiento donde las cajas negras deben en lo posible ser evitadas.

4. Validación

Con el propósito de realizar una prueba de concepto, los pasos identificados fueron implementados como una solución enmarcada en el proyecto ARIBEC [14]. El prototipo fue desarrollado en la plataforma Java y utilizó el API de Weka [15] como herramienta de minería. En este se procuró la implementación de patrones OO [5] de forma que se encapsularan los temas que actualmente se encuentran en investigación. Tras la ejecución de los escenarios de prueba se corroboró que el prototipo implementado reproducía el caso de estudio de forma semiautomática, lo cual facilita su ejecución a gran escala.

5. Conclusiones y trabajo futuro

Este artículo establece que el problema fundamental de generar metadatos extensionales mediante técnicas de minería de datos en el contexto de OV es un problema de escalabilidad. Para superar este reto el trabajo presenta un framework escalable basado en una estrategia inductiva. La estrategia parte de un caso de estudio particular para identificar la problemática relacionada con la caracterización y generación de metadatos extensionales; posteriormente se conceptualiza el conocimiento utilizado durante dicho caso para generalizarlo a otras fuentes de un mismo dominio. El framework aprovecha esta conceptualización al igual que el uso de técnicas de minería de datos.

Referencias

[1] Badillo., Julián. Arquitectura de servicios para la obtención y especificación de metadatos intencionales y extensionales en organizaciones virtuales a gran escala sobre plataformas Grid - Proyecto Aribec. Universidad de los Andes. Bogotá, Colombia. 2009.

[2] Berry M., Linoff G. Data Mining Techniques for Markerting, Sales and, Customer. 2004. Wiley Publising Inc.

[3] Camargo F., Arteta M. El sistema integral de información de la protección social - SISPRO. Bogotá, Colombia. 2006.

[4] Chapman P., Clinton J., Kerber R. CRISP-DM 1.0. Step-by-step data mining guide. Recuperado el 13 de 10 de 2008, de CRISP-DM 1.0: http://www.crisp-dm.org/CRISPWP-0800.pdf.

[5] Gamma E, Helm R., Johnson R., Vlissides J. Design Patterns: Elements of Reusable Object-Oriented Software. 1994. Addison-Wesley Professional Computing Series.

[6] Gómez V. Uso de minería de datos en la descripción de hospitalizaciones prolongadas. Proyecto de Grado Ingenieria de Sistemas y Computación Universidad de los Andes. Bogotá, Colombia. 2008.

[7] Grimm S., Hitzler P., Abecke A. Knowledge Representation and Ontologies. Semantic Web Services, págs. 51-106. 2008.

[8] Hand D., Mannila H., Smyth P. Principles of Data Mining. 2001. The MIT Press.

[9] Harrinson J. H. Introduction to the Mining Clinical Data. Clinincs in Laboratory Medicine , págs 1-7. 2008.

[10] Jain A.K., Murty M.N., Flynn P.J. Data Clustering: A Review. ACM Computing Surveys, 2000.

[11] Levy A., Rajaraman A, Ordille J. Querying Heterogeneous Information Sources Using Source Descriptions. Proceedings of the Twenty-second International Conference on Very Large Databases, págs. 251-262. 1996.

[12] Phillips J. and Buchanan B. Ontology-Guided Knowledge Discovery in Databases. Proceedings of the 1st international conference on Knowledge capture, págs 123-130. 2001.

[13] Zagoruiko N. G., Gulyaevskii S. E. and Kovalerchuk B. Ya. Ontology of the Data Mining Subject Domain, Pattern Recognition and Image Analysis, págs 349-356. 2007.

[14] Pomares A., Abásolo J., Roncancio C., Villamil P. Dynamic Source Selection in Large Scale Mediation Systems. Proceedings of the 1st international conference on Data Management in Grid and Peer-to-Peer Systems, págs 58-69 . 2008.

[15] The University of Waikato. Recuperado el 10 de 11 de 2008, de Weka 3: Data Mining Software in Java: http://www.cs.waikato.ac.nz/ml/weka.

Diego Ardila Álvarez. *Ingeniero de Sistemas y Computación, Universidad de los Andes.*

Natalia Valencia Lesmes. *Ingeniera de Sistemas y Computación, Universidad de los Andes.*

José Abásolo Prieto. *Ingeniero de Sistemas y Computación, Universidad de los Andes. D.E.A. en Informatique, INPG (Francia). Doctor de tercer ciclo en informática de las organizaciones, Universidad de Paris 9 Dauphine (Francia). Profesor titular Universidad de los Andes.*

María del Pilar Villamil. *Ingeniera de Sistemas y Computación, Universidad de los Andes. Magíster en Ingeniería de Sistemas y Computación, Universidad de los Andes. Docteur en Informatique, INPG (Francia). Profesora Asistente Universidad de los Andes.*