PERTINENCIA DE LA FORMALIZACIÓN DE DOMINIOS SEMI-FORMALMENTE DEFINIDOS EN EL ANÁLISIS INTELIGENTE DE DATOS

M.C. Karla Olmos Sánchez Ph.D. Jorge Rodas Osollo M.C. Luis Felipe Fernández

Centro de Ingeniería de Software

Universidad Autónoma de Ciudad Juárez

Resumen

Descubrir Conocimiento en Dominio Semi-Formalmente Definidos es un problema no trivial que generalmente requiere de soluciones de hechura a la medida, costosas y en las que se requiere invertir un tiempo considerable, cuando éstas son encontradas. Lo anterior debido a las grandes cantidades de conocimiento tácito o implícito que dificulta la organización del conocimiento en una estructura que ayude a transitar de un estado a otro hasta alcanzar una solución. Es así que el presente artículo somete a discusión la pertinencia de llevar a cabo una caracterización, quizás mediante la propuesta de una ontología, que ayude a los buscadores de conocimiento a tener mayor claridad de la situación que establezca la naturaleza del Problema que enfrentan y que les permitiría implementar una solución que produjera resultados más útiles y significativos y, posiblemente, con un menor esfuerzo.

Palabras clave: Análisis inteligente de datos, dominios semi-formalmente definidos.

Introducción

Debido a la tecnología digital, desde hace más de quince años está presente un crecimiento exponencial de datos; como consecuencia, se han creado técnicas y herramientas para su depuración y análisis cuya finalidad es obtener provecho de los datos que se van generando convirtiéndolos en información útil, es decir conocimiento. Tal conocimiento es producto del proceso que llevan a cabo las áreas de el Análisis Inteligente de Datos (IDA, Intelligent Data Analysis) y Descubrimiento de Conocimiento en Grandes Bases de Datos (KDD, Knowledge Discovery Data) (Fayyad et al, 1996). En ambas áreas se recurre a un proceso interactivo e iterativo de análisis de datos que involucra la preparación de los datos, la búsqueda de patrones, la evaluación y refinamiento de los patrones encontrados para determinar cuáles de ellos puedan ser considerados como nuevo conocimiento.

El KDD se ideo para trabajar con grandes cantidades de datos mientras que para el IDA esto no representa una restricción. Es así, que en lo general el KDD es utilizado para resolver Problemas

relacionados con grandes cantidades de datos y típicamente encontrados en ámbitos como: el financiero, varias ramas de ingeniería, seguridad informática, juegos, mercadotecnia… En el caso del IDA se ha utilizado para resolver Problemas Complejos e Imprecisos que por su naturaleza misma requieren de soluciones de hechura a la medida y que con mayor frecuencia se presentan en ciertos ámbitos que involucran Dominios como: medicina, biomedicina, educación, ingeniería de software, cambio climático… Una discusión detallada de las coincidencias y diferencias entre ellas se puede encontrar en (Lavrac *et al*, 2000).

Para que, tanto el KDD como el IDA brinden conocimiento es necesario tener claridad en cuanto a cómo es la relación e interactuación entre las partes comprendidas en el Dominio del Problema a resolver. Esto es un problema no trivial, especialmente en los Dominios que con más frecuencia atiende el IDA, debido a que, además de tratar con conocimiento explícito del Dominio específico del Problema a resolver, es necesario contemplar grandes cantidades de conocimiento tácito o implícito por lo que, generalmente, el conocimiento en estos Problemas carece de estructura o su estructura es incompleta y no es sencillo transitar entre estados para llegar a una solución.

Lo anterior, es bien conocido por quiénes nos dedicamos a obtener conocimiento y aunque solemos salir avantes de dicha situación, sería de gran utilidad contar con algún procedimiento que estableciera una estructura donde no la hay o al menos diera mayor certeza al cómo incluir o considerar el conocimiento tácito, de forma que se pueda mejorar la definición del problema, y el desarrollo o selección de algoritmos o metodologías con el objetivo de minimizar la cantidad de intentos fallidos al tratar de obtener conocimiento.

Es así, que el presente artículo somete a discusión la pertinencia de llevar a cabo una caracterización, quizás mediante la propuesta de una ontología, que ayude a los buscadores de conocimiento a tener mayor claridad de la situación que establezca la naturaleza del Problema que enfrentan y que les permitiría implementar una solución que produjera resultados más útiles y significativos y, posiblemente, con un menor esfuerzo.

La sección 2 revisa un conjunto amplio de referencias antecedentes de los Problemas en Dominios Parcialmente Definidos. La sección 3 resalta la Importancia del Conocimiento del Dominio

y la necesidad de establecer un instrumento para la caracterización de un Problema enmarcado por un Dominio Parcialmente Definido. En la sección 4 se mencionan algunas áreas de interés donde lo señalado en la sección anterior tendría un impacto directo. Finalmente en la sección 5 se presenta la Discusión y Trabajo Futuro.

Antecedentes de los Problemas de Dominios Parcialmente Definidos

Antecedentes

Un problema es algo desconocido que resulta de cualquier situación en la cual una persona busca completar una necesidad o alcanzar una meta. Para (Cao, 2006) cuando "…no existen formas propias de clarificar o refinar un problema, el proceso de solución de éste y los esfuerzos del proceso pueden ser extremadamente largos y costosos". Es así, que a lo largo de la historia se han realizado diversos esfuerzos por caracterizar los problemas. Por ejemplo, en el área de Inteligencia Artificial, Simon (Simon, 1973) postula que existen dos tipos de problemas: los Bien Estructurados y los Parcialmente Estructurados. Los primeros tienen una formulación correcta, se puede determinar el estado inicial y el estado meta a partir de esta formulación; y los operadores están bien definidos por lo que permiten progresar del estado inicial al estado meta sin complicaciones. Los Problemas Parcialmente Estructurados forman parte de una categoría residual y son todos los problemas que no cumplen con alguna característica de los Bien Estructurados.

Los problemas de matemáticas y ciencias generalmente se consideran como Bien Estructurados. Por otro lado, los problemas relacionados con ética, diseño, leyes y diagnóstico médico se consideran Parcialmente Estructurados.

Jonassen (1997) propone que los problemas se clasifiquen en Problemas Tipo Rompecabezas, Bien Estructurados y Parcialmente Estructurados. Para él "Los problemas parcialmente estructurados poseen múltiples soluciones y rutas de solución, pocos parámetros manipulables, y existe incertidumbre acerca de cuáles conceptos, reglas y principios son necesarios para la solución o cómo se organizan, y acerca de cuál solución es la mejor". Por otro lado, la diferencia de los Problemas de Acertijos con los Bien Estructurados es que los primeros son problemas descontextualizados, diseñados para manifestar procesos de pensamiento y razonamiento. Tales como el problema de las Torres de Hanoi y el acertijo de los Misioneros y Caníbales.

Rittel y Webber (1973) proponen otra clasificación de los problemas. Para ellos los problemas se clasifican en Rutinarios (tame) y Complicados (wicked). Los primeros son los relacionados a las matemáticas y al ajedrez (ejemplos similares que se mencionan para los Bien Estructurados).

Los Problemas Complicados son incompletos, contradictorios y sus requerimientos son variables. Las soluciones de estos problemas frecuentemente son difíciles de alcanzar e incluso de reconocer debido a la compleja interdependencia entre una gran cantidad de variables. El término Wicked Problem se ha utilizado recientemente en trabajos relacionados con los Sistemas de Soporte de Decisiones y Soft Computing (Zannier et al, 2007), (Klashner y Sabet, 2007) (Petkov et al, 2006).

Los Problemas se enmarcan en un Dominio, entendiendo éste último como un Universo del Discurso. Así como los problemas, los Dominios también se han intentado clasificar. Para Lynch y Aleven (2006) los Dominios típicamente connotan un área de estudio tal como la física, o un conjunto de problemas y hace un estudio exhaustivo de los Dominios Parcialmente Definidos. Para estos autores un Dominio Parcialmente Definido se caracteriza por 1) La falta de estándares para verificar la solución de los problemas 2) Las Teorías Formales en estos Dominios generalmente son consensuadas, típicamente usadas para guiar intuiciones y no para dictar resultados, 3) La Estructura de la Tarea es parcialmente definida y para resolver un problema se requiere determinar qué leyes o teorías se deben aplicar a la situación actual. 4) Los conceptos en estos dominios carecen de una definición absoluta, y 5) La división de los problemas en subproblemas no reduce la complejidad, ya que los subproblemas se restringen unos a otros, y ninguno de ellos puede ser resuelto sin considerar los efectos de los otros.

Sin embargo considera que los términos Parcialmente Estructurado y Parcialmente Definido son intercambiables y, para efectos de su trabajo, no establece una distinción entre los Problemas y Dominios.

Otros autores, sin embargo, han diferenciado entre Problemas y Dominios y han estudiado su relación. Por ejemplo, para Jonassen (1997) un problema tradicionalmente se define por un Dominio del Problema, un Tipo de Problema, un Proceso de Solución del Problema y una Solución. El Dominio del Problema consiste

del contenido (conceptos, reglas y principios) que definen los elementos del problema. En un trabajo posterior, Jonassen y Hung (2008) propone un modelo para la clasificación de problemas, donde la dificultad del problema se analiza y evalúa en términos de su naturaleza y nivel al examinar sus Dimensiones de Complejidad y Estructuración. La Dimensión de Complejidad se conforma de cuatro parámetros: la Amplitud del Conocimiento requerido para resolver el problema, el Nivel de Dificultad del Conocimiento del Dominio, la Complejidad de los Procedimientos de Resolución del problema, y la Complejidad Relacional. La Dimensión de Estructuración consiste de cinco parámetros: el desconocimiento del Espacio del Problema, la heterogeneidad de interpretaciones, la interdisciplinariedad, la dinámica del proceso de solución del problema y la cantidad de soluciones posibles.

Por otra parte en (Fournier-Viger et at, 2008) los autores consideran los Problemas Parcialmente Estructurados de acuerdo a Simon (1973) y argumenta que los Dominios que incluyen estos problemas y en los cuales los objetivos de enseñanza se enfocan en habilidades de resolución de problemas, se consideran Parcialmente Definidos.

En el trabajo de Mitrovic y Weerasinghe (2009), situado en el ámbito de los Tutoriales Inteligentes, se propone la necesidad de considerar dos dimensiones ortogonales cuando se trabaja con Problemas Indefinidos: el Dominio y la Tarea Instruccional. Los Dominios varían en términos de las Teorías del Dominio y pueden ser Bien o Parcialmente Definidos. Sin embargo, si el Dominio es Bien Definido no significa que las Tareas Instruccionales también lo sean. Como ejemplo, se considera la tarea de realizar un modelo de bases de datos de entidad relación. El Dominio es Bien Definido ya que los conceptos son claros y existe una sintaxis formal para elaborar los diagramas. Por otro lado, la Tarea es Parcialmente Definida ya que no hay un camino único para determinar la solución.

Por último, nos interesa resaltar los trabajas en el área de Descubrimiento de Conocimiento que han puesto especial atención a la caracterización de los problemas que trabajan. En particular el grupo Knowledge Engineering and Machine Learning Group (KEMLG) clasifica los problemas que atiende como Dominios Poco Estructurados (Gibert y Cortés, 2004)

y los caracteriza por: 1) Los elementos del dominio vienen descritos por conjuntos heterogéneos de variables, 2) Existe un conocimiento a priori adicional sobre la estructura del dominio, y 3) La complejidad inherente al dominio hace que el conocimiento que de él se tiene sea parcial (en este dominio existe gran cantidad de conocimiento implícito y grandes incógnitas) y no homogéneo (el grado de especifidad del conocimiento disponible es distinto para distintas partes del dominio).

De acuerdo a esta revisión referencial se puede observar que existe una necesidad de clasificación de los Problemas y que una forma de clasificarlos es de acuerdo a las características del Dominio en el que se enmarcan. Sin embargo, a pesar de las coincidencias, se puede observar una divergencia de opiniones. Lo anterior lo atribuimos a la visión que de estos Problemas y sus Dominios tienen los autores de acuerdo a su área de investigación. Por lo que es necesario establecer una postura que permita construir una argumentación.

Problemas Enmarcados en Dominios Semi-Formalmente Definidos

La figura 1 muestra un modelo conceptual como una propuesta para establecer una postura inicial acerca de los Problemas y los Dominios que toma como base la información referencial de la sección anterior (El modelo conceptual está creado de acuerdo a la sintaxis y semántica de los Diagramas de Clase de UML). Las líneas simples indican una relación de asociación, las líneas que finalizan en triángulos indican una relación de herencia y las líneas que finalizan en rombos rellenos indican una relación de agregación.

Para nosotros un Problema se compone de un Planteamiento, una o varias Soluciones y uno o varios Algoritmos de solución. El Planteamiento puede ser completo o incompleto. Cuando un Planteamiento no es completo consideramos que el Problema es Impreciso pero no necesariamente por su naturaleza. Por otro lado, la Solución tiene el atributo booleano de verificable dependiendo de si, como su nombre lo indica puede ser verificable o no. Aunque algunos autores proponen clasificar los problemas en más de dos categorías, nosotros consideramos que sólo es necesario dividirlos en dos: con Estructura y con Estructura Parcial. Los Problemas con Estructura Parcial pueden tener diversos niveles de “estructuración” y esto daría lugar a las diversas clasificaciones que proponen los autores. Un Problema se

asocia con un Conocimiento del Problema que sería todo el conocimiento que se tiene del Problema. Este Conocimiento se enmarca en un Dominio como Área de Conocimiento. Esta relación es uno a muchos ya que un Problema puede necesitar de diversas áreas para ser resuelto. El Dominio como área de conocimiento puede ser Formalmente Definido o Semi-Formalmente Definido. Definir los Dominios de esta forma y no de la generalmente encontrada en la literatura como Bien Definidos o Parcialmente Definidos obedece a la concepción de los Sistemas Formales. Más adelante se establece una definición a detalle de estos conceptos.

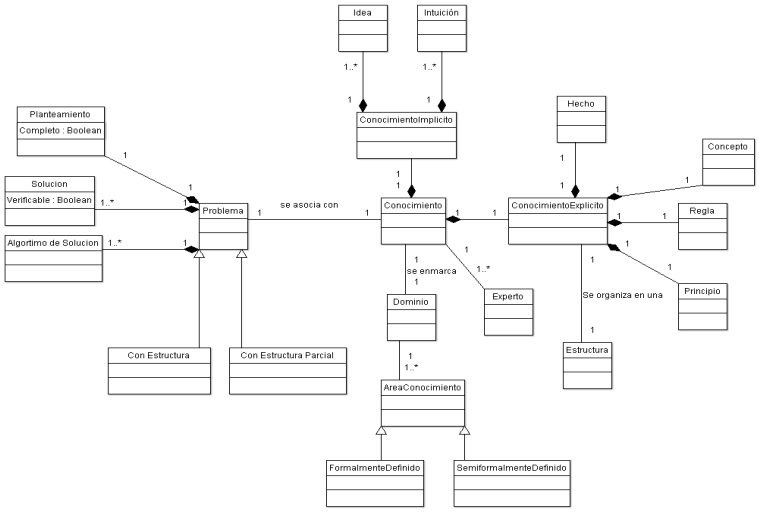

Figura 1. Modelo Conceptual Problemas y Dominios

El Conocimiento del Problema puede ser Explícito o Implícito. El primero se refiere a Hechos, Conceptos, Reglas y Principios, que puede ser representado de alguna forma tal que es posible compartirlo. El Conocimiento Implícito tiene que ver más con ideas e intuiciones, se adquiere por experiencia y no es susceptible de ser representado. Generalmente existen uno o varios Expertos que tienen el conocimiento o acceso a éste para resolver el Problema. Considerando que una Estructura es la forma en la cual algo se ordena u organiza, la estructura de un problema por lo tanto es la forma en que se organizan u

ordenan los Conceptos, Hechos, Reglas y Principios que nos permita un andamiaje para llegar a una solución. Depende del conocimiento explícito.

A continuación se establecen las definiciones para efectos de nuestro trabajo de Sistema Formal, Dominios Formalmente Definidos e Informalmente Definidos. Las definiciones propuestas recuperan gran parte de las ideas del trabajo de Lynch pero diferencia entre Dominio y Problema e incorpora explícitamente el conocimiento tácito como en el trabajo de Gibert.

Sistema Formal. Sistema axiomático compuesto de símbolos que se unen entre sí formando cadenas que a su vez pueden ser manipuladas según reglas para producir otras cadenas. El sistema formal es capaz de representar cierto aspecto de la realidad. Se llama formalización al acto de pretender capturar y abstraer la esencia de determinadas características del mundo real, en un modelo conceptual expresado en un determinado lenguaje formal. Un sistema así es la reducción de un lenguaje formalizado a meros símbolos, lenguaje formalizado y simbolizado sin contenido material alguno.

Dominios Formalmente Definidos. Una esfera de actividad, interés o función cuyos componentes explícitos (conceptos, hechos, reglas y principios) así como sus relaciones tienen una sintaxis y semántica bien definida y donde existen algoritmos determinísticos que garantizan soluciones finitas y verificables. Para resolver problemas en estos dominios generalmente no se requiere ninguna interpretación de la persona que resuelve el problema. El conocimiento para resolver problemas en estos dominios es en su mayoría explícito, por lo que se puede determinar la estructura de los problemas sin complicaciones. Algunos ejemplos de estos Dominios son la Física, Matemática, Química, etc.

Dominios Semi- Formalmente Definidos. Una esfera de actividad, interés o función cuyos componentes (conceptos, hechos, reglas y principios) así como sus relaciones tienen una sintaxis y semántica definidas de acuerdo a un consenso. Las soluciones de los problemas en estos dominios suelen ser diversas ya que dependen de la interpretación de la persona que resuelve el problema. Se puede construir una estructura parcial de los problemas con el conocimiento explícito que permita conducir a una

solución, aunque se requiere de grandes cantidades de conocimiento tácito para llegar a ella, el cual no es susceptible de representarse en la estructura. Algunos ejemplos de estos Dominios son la Medicina, la Ética, la Política, el área de Diseño en general como la Arquitectura o el Desarrollo de Software.

Por último, nos interesa resaltar la relación que existe en el Conocimiento Implícito y Explícito dependiendo de las características del Dominio, entre mayor definido se encuentre un Dominio mayor será la cantidad de conocimiento explícito que se tenga de los problemas y viceversa entre menos definido se encuentre el Dominio mayor será la cantidad de conocimiento implícito que se requiere para solucionarlos. Como puede apreciarse en la Figura 2.

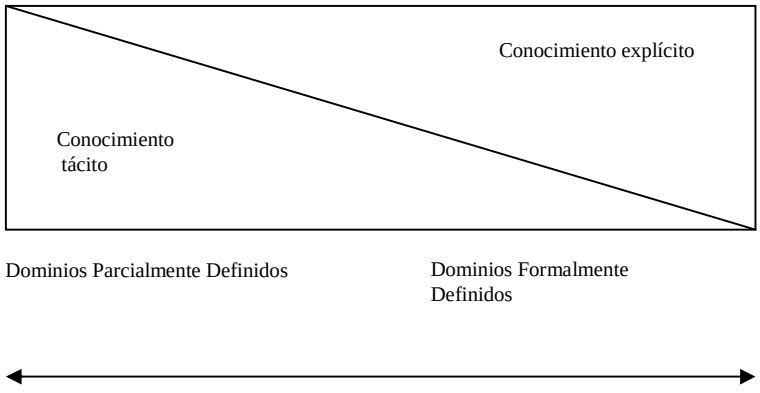

Figura 2. Relación entre Domino y Conocimiento

Necesidad de Caracterización de Dominios: Importancia del Conocimiento del Dominio

Una de las discusiones actuales de la comunidad de IDA tiene que ver con que los trabajos presentados en esta área se están enfocando más al análisis de algoritmos de minería de datos, cuando la esencia original de la disciplina era generar nuevo conocimiento para automatizar algunas de las habilidades de razonamiento de los analistas de datos.

Es así que una corriente actual del IDA es orientarse a resolver problemas prácticos de valor para la sociedad en ámbitos como: el cambio climático, la pérdida de hábitat, educación y medicina (Cohen y Addams, 2009). Según estos

autores para responder efectivamente a estos retos se deben replantear las siguientes actividades: el origen de los datos, los metadatos y la búsqueda de datos, el razonamiento acerca del contenido o el significado de los datos, las interfaces de usuario, la visualización de resultados e incluso considerar temas de privacidad y ética. Además, el conocimiento descubierto debe ser validado no sólo con mediciones técnicas sino también por el grado de valor que tiene para los expertos en el área de interés.

Como puede observarse, lo anterior implica incluir el Conocimiento del Dominio en el proceso de Descubrimiento de Conocimiento. El Conocimiento del Dominio se refiere al conocimiento que es válido y directamente utilizado en un Dominio preseleccionado de un desafío humano o una actividad de cómputo autónomo. Los especialistas y expertos utilizan y desarrollan su propio conocimiento del dominio. Siguiendo a (Viademonte y Burstein, 2006) el Conocimiento del Dominio puede ser clasificado en dos tipos, el conocimiento fáctico y el conocimiento relativo a la experiencia. El conocimiento fáctico consiste de conocimiento explícito del dominio, tal como hechos, datos, contexto y relaciones relevantes al problema de decisión; mientras que el conocimiento relativo a la experiencia consiste de conocimiento implícito del dominio que poseen los expertos.

El Conocimiento del Dominio incluye información acerca de las relaciones entre los objetos, tipos de atributos y otros aspectos semánticos; esto comprende el alcance de los valores y el significado de valores espaciales como los valores por defecto o las excepciones.

Diversos autores coinciden en la importancia del Conocimiento del Dominio como estrategia para mejorar los resultados del proceso de Descubrimiento de Conocimiento, en particular en Dominios cuyas características se asemejan a las de los Semi-Formalmente Definidos.

Por ejemplo, en (Redpath y Srinivasan, 2004) se argumenta que para automatizar un proceso de KDD con éxito se requiere capturar el Conocimiento del Dominio de tal forma que de soporte a los diferentes estados del proceso. Para estos autores algunos temas que requieren resolverse son a) El establecimiento una

clasificación general de Conocimiento del Dominio que pueda ser aplicable en diferentes dominios y b) La adopción de un lenguaje formal y notación que permita manipular las clases del Dominio del Conocimiento que sean reconocidas como estándares.

Por otro lado, Cao et al (2010) va más allá y sugiere la inclusión de la Inteligencia del Dominio en el proceso de minería de datos para salir avantes en el análisis de problemas de la vida real. La Inteligencia del Dominio consiste en el Dominio del Conocimiento y de los expertos, la consideración de restricciones, y el desarrollo de patrones difíciles de visualizar. Para Cao es el usuario o el experto, quien dice "si" o "no" a los resultados obtenidos.

Existen otras áreas como la biomedicina donde las características propias de los datos en estos dominios complican el proceso de extracción de conocimiento. Para enfrentar estos problemas, uno de las alternativas que se plantea es la necesidad de desarrollar métodos que sean capaces de utilizar alguna forma del conocimiento médico existente en las actividades de descubrimiento de conocimiento, ya que estas actividades sólo son significativas cuando consideran el conocimiento existente en el área de aplicación (Peek et al, 2009).

Por último, citando a Deng y Purvis (2009) "Sin el uso propio y suficiente del Conocimiento del Dominio en las aplicaciones de minería de datos se corre el riesgo de: a) elegir los algoritmos o modelos equivocados o sub-óptimos, b) malinterpretar los resultados del análisis de datos, y por lo tanto c) reducir la confianza del usuario en el uso de estos métodos".

Necesidad de Caracterización de Dominios

Como hemos mencionado anteriormente, actualmente existen grandes volúmenes de datos y una creciente necesidad de manipularlos para transformarlos en conocimiento. Sin embargo, mucho del trabajo en las áreas de KDD e IDA se ha enfocado en evaluar la eficiencia de los algoritmos con poco o nulo valor para las personas interesadas en la estructura de los datos, como el médico, el inversionista, el ambientalista, el ingeniero de software… Por tal motivo, diversos autores concuerdan que es necesario orientarse a resolver problemas de valor para la sociedad como el cambio

climático, la pérdida de hábitat, educación y medicina, entre otros.

Sin embargo, las técnicas o métodos convencionales para descubrir conocimiento en estos Dominios generalmente no generan resultados satisfactorios. La revisión referencial nos indica que en parte lo anterior es consecuencia de las características inherentes del Dominio al que pertenecen estos problemas. Lo que implica la utilización de grandes cantidades de conocimiento implícito para solucionarlos. Por lo que muchas ocasiones es preciso soluciones de hechura a la medida en las que se consume mucho tiempo para idearlas, ya que generalmente se realizan a prueba y error, y es probable que los métodos encontrados no funcionen para otros tipos de problemas, incluso con características similares.

A pesar de los diversos trabajos encontrados acerca de la importancia de considerar el Conocimiento del Dominio para extraer conocimiento, a excepción de los trabajos del grupo KEMLG, en el ámbito de Descubrimiento de Conocimiento, no se encontraron referencias dónde se mencionen explícitamente los término de Dominio Mal Estructurado, Dominio Mal Definido o algún concepto similar que evidencie que se intenta clasificar los Problemas de acuerdo al Dominio al que pertenecen.

Aunque no se descarta la idea que Problemas que pertenecen a estos Dominios están siendo analizados por otros grupos de investigación. Por ejemplo en el trabajo de Hassanien et al (2008) se analizan las relaciones entre variables psicosociales y niños con diabetes utilizando Rough Sets. Hassanien menciona que “el análisis de datos médicos frecuentemente concierne con el tratamiento de conocimiento incompleto, con el manejo de piezas inconsistentes de información y con la manipulación de varios niveles de representación de los datos”. Como puede observarse este problema pueda ser caracterizado como Semi-Formalmente Definido pero en el trabajo no se hace ninguna mención explícita del término como tal. Lo anterior puede ser debido a que no existe un parámetro que permita a los analistas identificar el tipo de Dominio que enmarca el Problema que atienden.

Aunado a esto, cabe mencionar que la definición propuesta en (Gibert y Cortés, 2004) Gibert y utilizada en

diversos trabajos del grupo KEMLG describe a grosso modo los Dominios Mal Estructurados y, aunque se ha utilizado con éxito en situaciones particulares (Gibert y Cortés, 2006), (Gibert y Pérez-Bonilla, 2005), (Rodas y Rojo, 2005), (Vazquez, 2008), consideramos que la enumeración de características es insuficiente para ayudar a los analistas de datos a definir el tipo de problema al que se enfrentan, lo cual les permitiría seleccionar los algoritmos o metodologías más apropiados para producir resultados más útiles y significativos y, posiblemente, minimizar la cantidad de intentos fallidos al tratar de obtener conocimiento.

Área de incidencia: Ingeniería de Software

Las áreas de incidencia de esta propuesta serían aquellas cuyos Dominios empaten con la definición propuesta de Semi-Formalmente Definidos como el diagnóstico médico, el cambio climático, educación, desarrollo de software… A continuación se analiza el impacto en Ingeniería de Software.

En Ingeniería de Software no sólo existe la necesidad de gestionar el conocimiento de la organización, sino que también es necesario entender el dominio para el cual el software será desarrollado. En este sentido Brooks (Brooks, 1987) señala que "… la dificultad del desarrollo de software es la especificación, diseño y prueba de sus constructos conceptuales y no la tarea de representar y probar la fidelidad de su representación".

Lo anterior implica que se debe poner especial cuidado en entender el dominio para definir debidamente los requerimientos del sistema para que el producto final se apegue en lo posible a las especificaciones del cliente. De igual forma, mucho del conocimiento de las organizaciones donde se implementaría el sistema de software, es conocimiento tácito difícil de describir y transformar en información que el analista pueda manipular (Friedrich y Var Der Poll, 2007).

En esta área, en (Deng y Purvis, 2009) se exponen un caso de estudio para estimar los esfuerzos de software en un proceso de desarrollo de software que combina el uso de algoritmos de minería de datos y Conocimiento del Dominio. Se propone una propuesta integral de minería de datos, donde la visualización, selección de características y modelos se conducen de acuerdo al Conocimiento del

Dominio. Este conocimiento también ayuda a validar el modelo de predicción así como asistir la interpretación de los resultados finales.

Dentro de la Ingeniería de Software existen otros Dominios como la Ingeniería de Requerimientos, en los que las características del problema como información de alta dimensión, dispersa y con errores, proveniente de expresiones cortas y ambiguas de los stakeholders, así como la necesidad de incluir a éstos en diferentes estados del proceso hace que las técnicas estándares de agrupamiento de minería de datos como K-means, agrupamiento aglomerativo jerárquico, y las técnicas probabilísticas no generen resultados satisfactorios (Duan, 2008). Descubrir conocimiento en estos Dominios requiere generalmente de soluciones de hechura a la medida que permita lidiar con la complejidad del problema en sí mismo y que incluyan el Conocimiento del Dominio.

Discusión

Una vez que se ha establecido el panorama general de los Dominios Semi-Formalmente Definidos y la necesidad de caracterizarlos, es importante discutir algunas cuestiones.

Primero, como puede observarse existe divergencia de opiniones y no hay mucho acuerdo en cuanto a la terminología utilizada en este tema y a las definiciones que dan soporte a esta temática. Por lo que es evidente la necesidad de una formalización de estos Dominios. La idea es transitar del Modelo Conceptual propuesto en la sección 2.2 a una Ontología formal que intente poner orden a las ideas de los diversos autores.

Otro punto es la necesidad de caracterizar los Problemas y Dominios desde el punto de vista del KDD o del IDA. Gran parte de las referencias encontradas acerca de los Problemas y Dominios se relacionan con el área educativa. Por ejemplo, Jonassen y Hung, (2008) caracterizan los Problemas para determinar que Dominios son factibles de ser enseñados utilizando el método de Aprendizaje Basado en Problemas. Sin embargo, hemos notado diferencias en el contexto del Descubrimiento de Conocimiento con la Educación, por lo que valdría la pena analizar este tema desde esta perspectiva. Por ejemplo, en el área educativa generalmente se enfocan en desarrollar habilidades para un Dominio como área de estudio en

particular, por lo que su interés es trabajar con Problemas de Dominio específicos.

En el área de Descubrimiento de Conocimiento generalmente el proceso es inverso, el Problema se genera de acuerdo a una necesidad de una o un grupo de personas interesadas en descubrir conocimiento en un Dominio en cual tienen cierta expertis. Otra diferencia es que cuando en el área educativa generalmente se enfocan a un Dominio, en el área de Descubrimiento de Conocimiento resolver un Problema puede necesitar conocimiento de diversos Dominios como área de estudio.

Por último, es importante hacer notar la necesidad de considerar explícitamente el conocimiento tácito en la definición de los Dominios Semi-Formalemente Definidos y analizar a detalle la forma en que afecta la caracterización de los Problemas y el grado en que se debe involucrar al experto en el Proceso de Descubrimiento de Conocimiento.

Trabajo Futuro

Por lo anterior anotado, es evidente que falta mucho trabajo por realizar. Sin embargo, consideramos que el siguiente paso es el desarrollo de una Ontología que proporcione una especificación explícita de la conceptualización de estos Dominios que permita poner orden a esta temática y que sea realizada desde el punto de vista del KDD y del IDA.

Una vez desarrollada la ontología se puede determinar de una manera más formal cuál es la relación específica entre los Problemas y los Dominios, además estudiar con mayor exactitud de qué forma las características del Dominio determinan las características del Problema.

Un trabajo a mediano plazo es explorar de qué forma las características del Dominio ayuden a seleccionar algoritmos o metodologías que minimicen el proceso de Descubrimiento de Conocimiento que permita enfocar los esfuerzos en solución de Problemas de valor para la humanidad y no solo en la evaluación de eficiencia de los diversos algoritmos.

Referencias

Brooks, F. 1987. No Silver Bullet - Essence and Accidentals of Software Engineerning. *IEEE Computing*, 20, 10 - 19.

Cao, L., Yu, P., Zhang, C., y Zhao, Y. 2010. *Domain Data Mining.* New York: Springer.

Cohen, P., y Addams, N. 2009. Intelligent Data analysis in the 21 st Century. Proceedings of the *8th International Symposium*

on Intelligent Data Analysis: Advances in Intelligent Data Analysis VIII. Lyon, France.

Deng, J., y Purvis, M. 2009. Software Effort Estimation: Harmonizing Algorithms and Domain Knowledge in an Integrated Data Mining Approach.

Duan, C. 2008. *Clustering and its Applications in Software Engineering*. DePaul University. USA: DePaul University.

Fayyad, U., Piatetsky-Shapiro, G., y Smith, P. (1996). *From Data Mining to Knowledge Discovery: an Overview*. The AI Magazine , 17 (3), 37-54.

Fournier-Viger, P., Nkambou, R., y Mephu Nguito, E. 2008. A Sequential Pattern Mining Algorithm for Extracting Partial Problem Spaces from Logged User Interactions. *3rd International Workshop on Intelligent Tutorial Systems in Ill-Defined Domains*, (págs. 46-55). Montreal Canada.

Friedrich, W., y Var Der Poll, J. 2007. Towards a Methodology to Elicit Tacit Dnowledge Domain Knowledge to Users. *Interdisciliplary Journal of Information, Knowledge and Managament*, 179-193.

Gibert, K., y Cortés, U. 2006. Clustering based on rules and Knowledge Discovery in ill-structured domains. *Computación y Sistemas*.

Gibert, K., y Cortés, U. 2004. Técnicas Híbridas de Inteligencia Artificial y Estadística para el Descubrimiento de Conocimiento y Minería de Datos. T*endencias de la Minería de Datos en España* , págs. 119-130.

Gibert, K., y Pérez-Bonilla, A. 2005. *Fuzzy box-plot based induction rules. Towards automatic generation of classes-interpretation.* EUSFLAT. Barcelona.

Hassanien, A., Abdelha, M., y Own, H. 2008. Rourh Sets Data Analysis in Knowledge Discovery: A Case of Kuwait Diabetic Children Patients. *Advances in Fuzzy Systems*, 8, 1-13.

Jonassen, D. H. 1997. Instructional design models for well-structured and III-structured problem-solving learning outcomes. *Educational Technology Research and Development*, 45 (1), 65-94.

Jonassen, D., y Hung, W. 2008. All Problems are not Equal: Implications for Problem-Based Learning. *The Interdisciplinary Journal of Problem-based Learning* , 2 (2), 6-28.

Klashner, R., y Sabet, S. 2007. A DSS Design Model for Complex Problems: Lessons from Mission Critical Infrastructure. *Decision Support Systems*, 43 (3), 990-1013.

Lavrac, N., Keravnou, E., y Zupan, B. 2000. Intelligent Data Analysis in Medicine. *Encyclopedia of Computer and Technology*, 9, 113-157.

Lynch, C., y Aleven, V. 2006. Defining "Ill-Defined Domains; A literature survey". *8th Conference on Intelligent Tutoring System*. Jhongli, Taiwan.

Mitrovic, A., y Weerasinghe, A. 2009. Revisiting Ill-Definedness and the Consequences for ITSs. *Frontiers in Artificial Intelligence and Applications* , 200, 375-382.

Peek, N., Combi, C., y Tucker, A. 2009. Biomedicarl *Data Mining. Journal Methods of Information in Medicine*, 48, 225-228.

Petkov, D., Petkova, O., Andrew, T., y Nepal, T. 2006. Mixing Multiple Criteria Decision Making with Soft Systems Thinking Techniques for Decision Support in Complex Situations. *Decision Support Systems*, 43 (4), 1615-1629.

Redpath, R., y Srinivasan, B. 2004. A Model for Domain Centered Knowledge Discovery in Database. Proceedings of the IEEE *4th International Conference on Intelligent Systems Designs and Applications*. Budapest, Hungary.

Rittel, H. W., y Webber, M. M. 1973. Dilemmas in a General Theory of Planning. *Policy Sciences*, 4 (2), 155-169.

Rodas, J., y Rojo, E. 2005. Knowledge Discovery in Repeated Very Short Serial Measurements with a Blocking Factor. Application to a Psychiatric Domain. *International Journal of Hybrid Intellignet Systems*, 2 (1/2005), 57-87.

Simon, H. 1973. *The Structure of Ill Structured Problems. Artificial Intelligence* , 4 (3-4), 181-201.

Vázquez, F. 2008. Caracterización e Interpretación Automática de Descripciones Conceptuales en Dominios Poco Estructurados. México, México: *Centro de Investigación en Computación del IPN*.

Viademonte, S., y Burstein, F. 2006. From knowledge discovery to computational intelligent: A Framework for Support Decision Systems. *En Intelligent Decision-making Support Systems Foundations, Applications and Challenges* (págs. 57-78).

Zannier, C., Chiasson, M., y Maurer, F. 2007. A Model of Design Decision Making Based on Empirical Results of Interviews with Software Design. *Information and Software Technology*, 49 (6), 637-653.