Representation of Knowledge from Intelligent Data Analysis as Requirements

Raymundo Camarena
al151096@alumnos.uacj.mx

Tania Olivier
al144857@alumnos.uacj.mx

Karla Olmos-Sánchez
kolmos@uacj.mx

Alexis Plata
al150666@alumnos.uacj.mx

Macario Ruiz-Grijalva
al150625@alumnos.uacj.mx

Jorge Rodas-Osollo
jorge.rodas@uacj.mx

Departamento de
Eléctrica y Computación
Instituto de Ingeniería y
Tecnología, UACJ
Maestría en Cómputo
Aplicado
Cd. Juárez, Chih., México

ABSTRACT

The technological dependence that society suffers has propitiated that the information stored in servers of the whole world has an exponential growth. As a consequence of this phenomenon the useful information consultation becomes complex and it is necessary to use intelligent methods to reach it. This article mentions how Intelligent Data Analysis helps to search for knowledge within information banks regardless of their size.

RESUMEN

Título: Representación del Conocimiento proveniente del Análisis Inteligente de Datos como Requisitos

La dependencia tecnológica que sufre la sociedad ha propiciado que la información almacenada en servidores de todo el mundo tenga un crecimiento exponencial. Como consecuencia de este fenómeno la consulta de información útil se torna compleja y es necesario emplear metodos inteligentes para llegar a ella. Este artículo menciona como el Análisis Inteligente de datos ayuda a la búsqueda de conocimiento dentro de bancos de información sin importar su tamaño.

Palabras Clave

Análisis inteligente de datos; Representación del Conocimiento; Requisitos

1. INTRODUCCIÓN

La transferencia y el almacenamiento diario de i nformación digital crece exponencialmente. Esto ha obligado a l os i nvestigadores a desarrollar técnicas y herramientas para l a depuración y el análisis de l a i nformación con l a finalidad de identificar i nformación útil, es decir conocimiento.
El conocimiento adquirido de la información puede ser producto del proceso llamado Análisis Inteligente de Datos (IDA, Intelligent Data Analysis) o del Descubrimiento de Conocimiento en Grandes Bases de Datos (KDD, Knowledge Discovery Data) [1].
En ambas áreas se recurre a un proceso i nteractivo e i terativo de análisis que i nvolucra l os siguientes pasos: conocimiento del dominio, preparación de l os datos, extraer i rregularidades, búsqueda de patrones ocultos, evaluación y refinamiento de l os patrones encontrados para determinar cuáles de ellos puedan ser considerados como nuevo conocimiento.
El KDD se i deó para trabajar con grandes cantidades de datos, regularmente está asociado con el análisis de i nforma-ción para el área financiera, i ngeniería, seguridad i nformáti-ca, El IDA, al no tener restricción de la cantidad de datos para analizar fue diseñado para dar soluciones a problemas complejos e imprecisos mediante el análisis de la información relacionada con el problema. Regularmente el IDA es asociado en el análisis de la información médica, biomédica, educación, ingeniería de software, cambio climático, etcétera. Una discusión detallada de las coincidencias y diferencias entre ellas se puede encontrar en [2].
Para que tanto el KDD como el IDA brinden conocimiento, es necesario tener claridad en cuanto a cómo es la relación e interacción 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, ya que además de tratar con conocimiento explícito del dominio del problema a resolver, es necesario contemplar grandes cantidades de conocimiento tácito o implícito que por su naturaleza carece de estructura, siendo más complejo el análisis de la información.

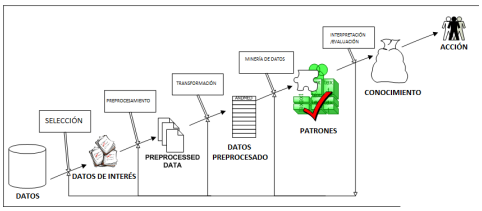

Figura 1: Esquema representativo del proceso general del KD. El KD consta de seis etapas principales que parte de los datos y concluye en el descubrimiento de conocimiento para la toma de desiciones

La clave para obtener conocimiento al aplicar el IDA es la habilidad para reconocer el problema a resolver. Algunas de las técnicas empleadas son: agrupación, visualización de datos, interpretaciones de datos en el tiempo.
Un ejemplo de la aplicación del IDA en medicina mencionado por [3], es la implementación de los siguientes métodos: Reglas simbólicas explicitas para entrenar casos y usarlos en inducción basada en reglas o en arboles de decisiones, clasificando los datos de interés. El aprendizaje basado en

instanciación almacena casos ya entrenados para referenciar, cuándo hay nuevos casos se clasifican comparandolos con los casos almacenados (CBK).
Otro caso es la probabilidad condicional empleando la abstracción de datos en el tiempo para el ámbito médico donde está relacionado con el proceso que envuelve al seguimiento de enfermedades para trabajar con un historial médico. Por lo general, en la mayoría de los modelos el tiempo se representa como una integral y permite optimizar la formación de posibles soluciones.
Las abstracciones temporales sobre el tiempo y los datos juegan un rol fundamental para los sistemas médicos basados en conocimiento ya que es necesario aplicar el conocimiento en datos específicos de pacientes con el objetivo de emitir un pre diagnóstico que apoye la decisión de alguna terapia para el paciente, también el monitoreo de terapias y acciones complementarias.
Los sistemas inteligentes agrupan algoritmos que implementan o emulan distintos modelos de aprendizaje, comportamientos de ciertos sistemas biológicos, entre otras cosas, y su aplicación a la resolución de problemas complejos. Entre los problemas abordados en este campo está el de inducir conocimientos a partir de datos o ejemplos, esto resulta una alternativa de solución a problemas que no pueden ser resueltos mediante métodos convencionales, tales como métodos estadísticos, modelos matemáticos, entre otros.
Todos estos métodos convencionales son esencialmente matemáticamente formales. En contraposición, los métodos basados en sistemas inteligentes están orientados principalmente hacia el desarrollo de descripciones simbólicas de los datos, que puedan caracterizar uno o más grupos de conceptos, diferenciar distintas clases, seleccionar los atributos más representativos de grupos de datos, ser capaces de predecir secuencias, etc. Estos métodos son esencialmente cualitativos, lo cual permite el descubrimiento de patrones en estructuras de información [3].

Análisis inteligente de datos para la búsqueda de conocimiento

En el presente artículo se somete a discusión la pertinencia de llevar a cabo una caracterización que apoye a los buscadores de conocimiento a arrojar información certera para la identificación del dominio de la solución del problema teniendo resultados útiles, significativos y eficientes. 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 de las representaciones de conocimientos empleadas en el IDA. En la sección 5 se indican algunas de las áreas de interés donde lo señalado en la sección 3 tendría un impacto directo. Finalmente en la sección 6 se presenta la Discusión y Trabajo Futuro.

2. ANTECEDENTES DE LOS PROBLEMAS DE DOMINIOS PARCIALMENTE DEFINIDOS

A lo largo de la historia se han realizado diversos esfuerzos por caracterizar los problemas. Por ejemplo, en el área de Inteligencia Artificial, Simon [4] 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.
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 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 Lynch 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 sub problemas no reduce la complejidad, ya que los sub problemas 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 [5].
En el trabajo de Gibert [6] se clasifican los problemas que se atienden como Dominios Poco Estructurados 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 especificar el 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.

3. CONOCIMIENTO DEL DOMINIO Y NECESIDAD DE CARACTERIZAR EL PROBLEMA QUE ENMARCA

La necesidad de caracterizar problemas pertenecientes a dominios de estructura informal representa un problema no

trivial para el proceso de extracción de conocimientos. En las dos subsecciones posteriores se abordarán algunos puntos importantes relacionados al IDA y su implementación en dominios de estructura informal.

3.1 Importancia del Conocimiento del Dominio

Una de las discusiones actuales de la comunidad de IDA tiene que ver con que los trabajos en esta área se enfocan 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 [7]. Según Cohen 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. Un esquema general del proceso del IDA que engloba sus características es presentado en la figura 2. 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.

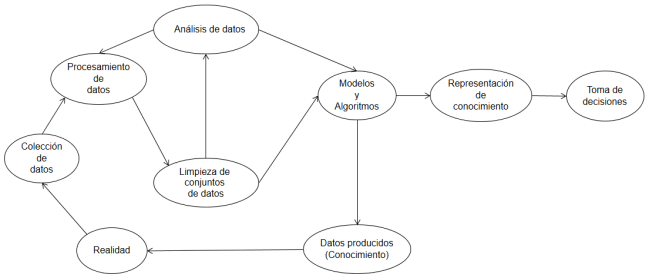

Figura 2: Esquema representativo del proceso general de aplicación del IDA para la toma de desiciones inteligentes

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 [8] 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 [9] 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 Redpath 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 [10] va más allá y sugiere la inclusión de la Inteligencia del Dominio en el proceso de minería de datos para salir avante 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 [11].
Por último, citando a Deng [12] “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”.

3.2 Necesidad de caracterización de Problemas enmarcados en Dominios de Estructura Informal

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.

4. REPRESENTACIÓN DEL CONOCIMIENTO

Una de las metas del Análisis Inteligente de Datos consiste en generar conocimiento nuevo auxiliándose de representaciones de conocimiento. Para cumplirla es necesario pasar por un proceso donde los datos y la información se colectan, clasifican, organizan e integran para que puedan agregar valor resultando en conocimiento nuevo. También es primordial adaptar la representación del conocimiento de acuerdo a las necesidades del problema y visualizar las especificaciones del problema o necesidad como requisitos. Uno de los principales retos es encontrar una representación de conocimiento adecuada para garantizar que el conocimiento representado comunique lo correcto y facilitar su interpretación. Una manera de representación de conocimiento es a través de los requisitos que consiste básicamente en la abstracción de necesidades o condiciones a satisfacer de un problema dado. Los requisitos deben tener como característica principal que sean medibles, comprobables, sin ambigüedades ni contradicciones. Existe una variedad de opciones para crear representaciones de conocimientos en la ingeniería de requisitos. Estas representaciones pueden ser tan sencillas como una lista o ser presentadas tan completas como lo son las Ontologías. Las matrices, los modelos, los métodos, los algoritmos o en su caso alguna combinación de estos pueden ser representaciones validas en la ingeniería de requisitos. El escoger que tipo de representación a usar dependerá en su totalidad del tipo de problema que se aborda y encontrar la que facilite su interpretación. La manera en que se comunique o se trate de transmitir conocimiento es amplia y existen esfuerzos para que sea de manera formal. Para cada área de aplicación ha evolucionado a distintos ritmos. Actualmente las distintas ramas de conocimiento han desarrollado y formalizado sus técnicas para poder representar el conocimiento. El objetivo fundamental de la representación del conocimiento es facilitar la inferencia, el sacar conclusiones a partir de dicha representación.

5. ÁREA DE INCIDENCIA: INGENIERÍA DE REQUISITOS

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 [13] 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 es conocimiento tácito difícil de describir y transformar en información que el analista pueda manipular para resolver algún problema o satisfacer alguna necesidad [14].

En la aplicación de la Ingeniería de Requisitos para procesos de desarrollo de software existen otros Dominios en los que las características del problema hace que las técnicas estándares de agrupamiento de minería de datos como *K-means*, agrupamiento acumulativo jerárquico, y las técnicas probabilísticas no generen resultados satisfactorios [15]. 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.

6. DISCUSIÓN Y TRABAJO FUTURO

6.1 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. Generalmente se enfocan en desarrollar habilidades para un Dominio como área de estudio en particular, por lo que su interés es trabajar con Problemas específicos de un Dominio particular. 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 del que se considera especialista. 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 formalmente 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.

6.2 Trabajo futuro

Se detectaron las siguientes necesidades:

1. 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.

2. 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 determinar con mayor exactitud de qué forma las ca-

racterísticas del Dominio determina las características del Problema.

3. 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.

7. REFERENCIAS

[1] U.˜Fayyad, G.˜PiatetskyShapiro, and P.˜Smith, "From data mining to knowledge discovery: an overview," *The AI Magazine*, vol.˜17, no.˜3, pp.˜37–54, 1996.

[2] N.˜Lavrac, E.˜Keravnou, and B.˜Zupan, "Intelligent data analysis in medicine," *Encyclopedia of Computer and Technology*, vol.˜9, pp.˜113–157, 2000.

[3] N.˜Lavrac, E.˜Keravnou, and B.˜Zupan, *Intelligent Data Analysis in Medicine and Pharmacology.* New York: Springer science and Business Media, LLC, 1997.

[4] H.˜Simon, "The structure of ill structured problems," *Artificial Intelligence*, vol.˜4, no.˜3-4, pp.˜181–201, 1973.

[5] C.˜Lynch and V.˜Aleven, "Defining ill defined domains a literature survey," *8th Conference on Intelligent Tutoring System*, 2009.

[6] K.˜Gibert and U.˜Córtes, "Técnicas híbridas de inteligencia artificial y estadística para el descubrimiento de conocimiento y minería de datos," *Tendencias de la Minería de Datos en España*, pp.˜119–130, 2004.

[7] P.˜Cohen and N.˜Addams, "Intelligent data analysis in the 21 st century," in *Proceedings of the 8th International Symposium on Intelligent Data Analysis: Advances in Intelligent Data Analysis VIII.*, (Lyon, France), 2009.

[8] S.˜Viademont and F.˜Burstein, "From knowledge discovery to computational intelligent: A framework for support decision systems," *En Intelligent Decision-making Support Systems Foundations, Applications and Challenges*, pp.˜57–78, 2006.

[9] R.˜Redpath and B.˜Srinivasan, "A model for domain centered knowledge discovery in database," in *Proceedings of the IEEE 4th International Conference on Intelligent Systems Designs and Applications*, (Budapest, Hungary), 2004.

[10] L.˜Cao, P.˜Yu, C.˜Zhang, and Y.˜Zhao in *Domain Data Mining*, (New York), Springer, 2010.

[11] N.˜Peek, C.˜Combi, and A.˜Tucker, "Biomedicarl data mining," *Journal Methods of Information in Medicine*, vol.˜48, pp.˜225–228, 2009.

[12] J.˜Deng and M.˜Purvis, "Software effort estimation: Harmonizing algorithms and domain knowledge in an integrated data mining approach.," 2009.

[13] F.˜Brooks, "No silver bullet - essence and accidentals of software engineerning," *IEEE Computing*, vol.˜20, pp.˜10–19, 1987.

[14] W.˜Friedrich and J.˜V.˜D. Poll, "Towards a methodology to elicit tacit dnowledge domain knowledge to users," *Interdisciliplary Journal of Information, Knowledge and Managament*, pp.˜179–193, 2007.

[15] C.˜Duan, "Clustering and its applications in software engineering," *DePaul University. USA*, 2008.