Línea 1: "Modelado semántico de datos, su captura y evolución: desde los requisitos hasta el sistema de software"

Director de Línea: RIVERO, Laura Celia
Personal Participante: FERRAGGINE, Viviana Elizabeth
MASSA, José María
RIDAO, Marcela Natividad
TRISTAN, Paula Mariela

OBJETIVOS Y/O FINALIDAD

Asistir al proceso de desarrollo de software, facilitando la especificación de los requisitos y contribuyendo al proceso de obtención de un modelo conceptual de datos incluyendo nuevas simbologías, que permitan plasmar reglas de integridad no consideradas por los algoritmos clásicos. Incorporar al modelado de datos aspectos relacionados con información no convencional y con el manejo de grandes volúmenes de datos.

Objetivos específicos

- Proponer una especificación exhaustiva de componentes constructores de los modelos de datos de aquéllos aspectos ignorados o insatisfactoriamente definidos en los modelos convencionales siguiendo en un enfoque objeto-relacional,
- Definir los algoritmos de conversión de los componentes especificados para la obtención de un Esquema Lógico Estándar (ELE) de datos que redunde en simplicidad y completitud y del ELE a un esquema físico en SQL4.
- Definir un proceso de reingeniería sin pérdidas, desde el ELE hacia el modelo conceptual de datos de partida.
- Extender y aplicar los conceptos de integridad semántica a datos no convencionales relativos a análisis morfométricos, geométricos y geográficos, con énfasis en recursos geométrico-computacionales de herramientas open source.
- Contribuir a la definición del esquema conceptual de grandes volúmenes de datos segmentados por necesidades de procesamiento, comunicaciones y espacio de almacenamiento (sharding y big data estructurados)
- Incorporar a herramientas de captura de datos basados en Sistemas de Información Geográfica (SIG) un conjunto de mecanismos que soporten patrones de consultas para extraer información enriquecida desde grandes volúmenes de información.
- Incorporar mecanismos de integridad a herramientas para la generación automática de aplicaciones de captura de datos de diferentes orígenes basados en SIG.
- Aplicar técnicas cuantitativas o combinaciones de técnicas cualitativas y cuantitativas en diferentes etapas del proceso de requisitos y lograr información útil para las posteriores etapas del desarrollo de software. En particular, se analizarán los modelos de LEL (Léxico Extendido del Lenguaje), Escenarios Actuales y Futuros, y el modelo de especificación de Requisitos, sistematizando el metaconocimiento disponible en ellos, con el fin de su reuso semántico. En este contexto, se propone:
  - Mejorar los algoritmos dirigidos por fuerzas utilizados en la detección de agrupamientos en el modelo del LEL, aplicando diferentes sistemas de fuerzas, extendiendo el algoritmo a 3 o 4 dimensiones, ponderando los vínculos entre nodos según diferentes criterios, etc.
  - Extender la aplicación de algoritmos para la detección de clusters semánticos a otros modelos de la Ingeniería de Requisitos, como Escenarios Actuales, Escenarios Futuros y documento de especificación de Requisitos.
  - Proponer nuevos algoritmos para la detección de agrupamientos en el modelo del LEL, aplicando técnicas de visualización de grafos que modelen vínculos semánticos, y otras técnicas de clusterización.
- Proponer mecanismos de visualización de los grafos utilizados para representar modelos del proceso de Ingeniería de Requisitos, incorporando interacciones que ofrezcan mayor información y permitan una mejor comprensión del modelo visualizado, y los eventuales agrupamientos detectados.
- Continuar con el desarrollo de transformaciones ATL aplicadas a modelos de requisitos, incorporando otras reglas que mejoren el modelo conceptual obtenido.

ESTADO ACTUAL DEL CONOCIMIENTO

El proceso que va desde la captura e identificación de las necesidades de los clientes y usuarios, pasando por la adquisición de las propiedades de los datos, capturados como requisitos en el Universo de Discurso (UdeD), hasta su definición en un sistema de procesamiento se ha tornado cada vez más complejo, enfrentando, entre otras dificultades, problemas como la ambigüedad, siempre presente en el mundo real, la completitud como una meta la mayoría de las veces inalcanzable. Es así que la captura e interpretación de las Reglas del Negocio (RN) de los datos, su representación y su rol en el mantenimiento eficiente de la integridad de los datos, son etapas del desarrollo de software imprescindibles y altamente significativas.

Un buen diseño de base de datos, que es soporte para cualquier sistema de información, es crucial para obtener una base robusta y consistente. En consecuencia, para lograr un diseño completo y exacto, se necesitan metodologías que permitan representar todos los aspectos relevantes del UdeD (Badía y Lemire, 2011). Si bien las RN han sido ampliamente tratadas por diversas metodologías en ingeniería de la información, siguen teniendo aspectos insuficientemente explorados respecto de su elicitación desde el UdeD, su formalización e implantación en un modelo de datos y posteriormente su materialización final en una base de datos.

Los **modelos de datos** más difundidos en la actualidad, Modelo de Entidades y Relaciones Extendido (MERExt) o Unified Modeling Language (UML), proveen lenguajes de especificación y notación todavía carentes de la fuerza expresiva necesaria para representar características valiosas del UdeD. Esta limitación tiene al menos dos consecuencias: a) la especificación de

algunas restricciones de integridad que podrían ser representadas adecuadamente con las estructuras disponibles se aplazan infundadamente hasta últimas etapas del proceso de diseño de datos; b) surgen propuestas que extienden los modelos mencionados con conceptos 'foráneos', ofreciendo mecanismos de abstracción adicionales.

Muchos autores han escrito acerca de las reglas de transformación del modelo conceptual de datos al esquema lógico estándar (ELE), proponiendo algoritmos que actualmente son de uso masivo, pero estos resultan incompletos en algunos aspectos, ya que existen ambigüedades en ciertas transformaciones y se percibe una pérdida de ortogonalidad y carencia de una definición formal (Chen, 1976; Pieris, 2013a; Pieris y Rajapakse, 2012; Cuadra y otros, 2012).

En Pieris y Rajapakse, 2012 y Pieris, 2013 se presenta un esbozo de una innovadora notación del modelo conceptual de datos al ELE y un conjunto de reglas de construcción para un modelo de datos objeto-relacional basado en el MERExt, acompañada por extensiones para su conversión en un ELE (Pieris, 2013b). Sin embargo, algunos aspectos del algoritmo extendido propuesto no han sido suficientemente explorados,

Desde la perspectiva de los **volúmenes de datos** que los sistemas de información actuales manejan, existe una amplia diversidad y durante un cuarto de siglo, la base de datos relacional ha sido el modelo dominante para la gestión de datos convencionales y de volúmenes manejables. Hoy en día las bases de datos "NoSQL" están cobrando protagonismo como un modelo alternativo, ya que están fuertemente asociadas a Big Data, nombre que se le ha dado a un cúmulo de conceptos asociados con volúmenes gigantescos de datos y datos convencionales y no convencionales. Esta nueva visión, ha planteado grandes desafíos en lo que respecta a modelos de datos orientados a Big Data. Esto significa la necesidad de particionar o segmentar una base de datos en estructuras particulares bien definidas. El sharding (segmentación) es una técnica que consiste en dividir desde una simple tabla hasta una base de datos 'horizontalmente', de manera que pueda mejorarse la escalabilidad. Típicamente estos shards o segmentos están localizados en tablas en diferentes bases de datos y hasta en diferentes localizaciones físicas. En este contexto uno de los aspectos que resalta es el problema de mantener la integridad de los datos, ahora dispersos

Desde el punto de vista de los datos, actualmente son provenientes de nuevas y diversas fuentes de información, constituyéndose en un campo de investigación en constante evolución, debido a sus diferentes cualidades. Específicamente los Sistemas de Información Geográfica, Médica, Documental, o Morfométrica, plantean nuevas perspectivas y necesidades de procesamiento, que difieren de aquéllas ya firmemente establecidas en los sistemas que procesan datos convencionales. Actualmente existe un gran número de especialistas desarrollando aplicaciones de software que necesitan la incorporación y tratamiento de nuevas fuentes de datos, la mayoría de las cuales se basan estrictamente en desarrollos científicos tecnológicos. Así, el SIG permitiría encarar diversos problemas actuales del tratamiento de información, proyecciones temporales, calidad semántica y otros, para obtener información enriquecida.

Otro tipo de información no convencional que requiere tratamiento específico, se refiere a las características morfológicas de una población, en forma de configuraciones de puntos anatómicos (landmarks). Se entiende que estas configuraciones pueden ser manipuladas mediante adaptaciones de métodos conocidos de administración de otros datos pasibles de tratamiento geométrico (Torcida et al 2014).

Mas aún, con el advenimiento de los sistemas de información de propósitos específicos, se suma a lo anterior el desarrollo de una diversidad de aplicaciones sobre dominios que se entrelazan y cooperan en cada desarrollo de software (sistemas de información médica, geográfica, documental, morfométrica, etc.) planteando nuevos desafíos que surgen por la naturaleza especial de los datos que manejan.

Por otra parte, los requisitos evolucionan, y esa evolución debe ser considerada como un aspecto de gran importancia en el desarrollo de cualquier tipo de sistema de información. Dicha característica de los requisitos ha sido estudiada por muchos autores y, en particular, es uno de los pilares del modelo Requirement Baseline (Leite et al., 1997). Éste modelo involucra una estrategia dirigida por modelos basados en lenguaje natural, que produce un glosario del UdeD, el Léxico Extendido del Lenguaje (LEL) y aplica la técnica de escenarios con el propósito de obtener conocimiento acerca del problema y capturar los requisitos del software (Leite et al 2000; 2004). Los modelos de LEL y escenarios no sólo sirven para registrar información, sino que motivan su elicitación. Son en sí mismos promotores de la captación de información, y facilitan la transmisión de información por parte de los clientes y usuarios. (Hadad 2007; 2009).

Los modelos mencionados, del mismo modo que ocurre con el resto de los modelos construidos a lo largo del proceso de desarrollo de software, son creados con propósitos y estructuras bien definidas. Estas estructuras han sido concebidas para maximizar la expresividad de cada modelo en relación con su propósito. Pese a esto y muy posiblemente debido a esto, puede ocurrir que en los mismos exista información no perceptible durante su uso rutinario. En el caso de los modelos de IR basados en lenguaje natural, se ha observado que una segunda lectura de algunos de ellos permite la adquisición de al menos parte de esa información oculta.

Tradicionalmente, la casi totalidad de los métodos y de las innovaciones en el área de la Ingeniería de Requisitos (IR) se han basado en enfoques cualitativos guiados por la problemática estudiada y/o por los modelos involucrados (Goguen 1993; Karlsonn 1997; Mylopoulos 1999; van Lamsweerde 2001). En este trabajo se pretende aplicar técnicas cuantitativas, o combinaciones de técnicas cualitativas y cuantitativas en diferentes etapas del proceso de requisitos y lograr información útil para las posteriores etapas del desarrollo de software que no se evidencia directamente durante dicho proceso.

Se propone aplicar estrategias cuantitativas o mixtas que revisen, desde puntos de vista semánticamente diferentes, los modelos del proceso de IR antes mencionados. La información semántica obtenida a partir de la aplicación de las estrategias propuestas será útil en etapas posteriores del desarrollo de software, y retroalimentará, cuando corresponda, las heurísticas ya conocidas de construcción de LEL y Escenarios

A continuación, se describen brevemente algunas de las técnicas utilizadas hasta el momento.

Detección de Agrupamientos Semánticos mediante algoritmos dirigidos por fuerzas.

Existen numerosos ejemplos, en diversas disciplinas, donde la detección de agrupamientos representa una contribución significativa a la mejor comprensión del fenómeno que está siendo estudiado (Das et al. 2011, Mo et al. 2012, Zimmermann et al. 2012). En particular, la detección de agrupamientos es utilizada con diversos fines en la Ingeniería de Requisitos. Por ejemplo, en (Duan 2009) los requisitos se agrupan en clusters, con el objetivo de priorizarlos.

Algunos de los modelos de la Ingeniería de Requisitos pueden ser estudiados desde el punto de vista estructural. En particular, uno de los más promisorios es el LEL. Si se observa un LEL bajo la óptica estructural se puede construir un grafo donde los símbolos sean los nodos y las menciones a otros símbolos sean arcos dirigidos, visualizándolo como una suerte de red lingüística con una estructura claramente compleja. Esta forma de representación permite observar que, además de la información explícita almacenada en cada nodo, existe información implícita empotrada en la estructura de las relaciones entre los nodos. Uno de los posibles resultados del estudio del conocimiento empotrado en la estructura del LEL está relacionado con la existencia o no de agrupamientos de símbolos por alguna razón no siempre visibles en la visualización plana del grafo o en su navegación interactiva. Estos agrupamientos o sub-agrupamientos, de existir, permiten la visualización de componentes de posibles taxonomías del proceso del negocio.

Se han visualizado los grafos correspondientes al LEL de diferentes casos de estudio mediante algoritmos dirigidos por fuerzas (Eades, 1984; Walshaw, 2003; Aiello, 2004). En estos algoritmos, los nodos conectados se atraen entre sí, y los nodos no conectados se repelen.

En trabajos previos (Ridao y Doorn, 2008; 2011), y durante la ejecución del proyecto 03/C248, se ha podido comprobar, mediante la aplicación de este tipo de algoritmos la presencia o ausencia de agrupamientos en los grafos representando el LEL para diferentes casos de estudio. Para casos donde un análisis semántico previo descartó la idea de agrupamientos, el algoritmo lo comprobó, mientras que, para casos donde se conocía previamente la existencia de agrupamientos, el algoritmo los detectó claramente. Más aún, el algoritmo permitió detectar subdivisiones existentes no observadas previamente. (Ridao y Doorn, 2013). Se estudió la aplicación de diferentes conjuntos de fuerzas para la atracción y repulsión de los nodos en el algoritmo, obteniendo resultados que permitieron visualizar los agrupamientos con mucha mayor claridad. Se plantearon también métricas que brindan una base para determinar el número de clusters de acuerdo a la forma geométrica de los grafos resultantes a partir de la aplicación del algoritmo (Ridao y Doorn, 2014; Ridao y Doorn, en prensa).

Detección de Agrupamientos Semánticos: Otras técnicas.

Continuando con el objetivo de buscar agrupamientos semánticos, se aplicó la técnica de clusterización de documentos a algunos de los casos de estudio ya analizados con los algoritmos antes descriptos. Estas técnicas de Minería de Textos, detectan agrupamientos de documentos por la repetición de términos o sus derivados en los mismos. Como resultado, se detectaron los mismos grupos para un caso donde los algoritmos dirigidos por fuerzas habían arrojado los resultados más claros. Sin embargo, para el resto de los casos estudiados hasta el momento, los resultados no fueron tan concluyentes. Se pretende continuar estudiando la aplicación de esta técnica, con el fin de mejorar dichos resultados.

Estimación del número de requisitos no detectados

Los métodos y herramientas de la Ingeniería de requisitos intentan capturar y especificar los requisitos del software, maximizando calidad y completitud. Sin embargo, el problema de la completitud es una amenaza constante en la calidad. Este problema, es similar a lo que ocurre en otras áreas del conocimiento. Otis (1978) introdujo un método para estimar el tamaño de una población cerrada de animales basándose en los datos de captura y recaptura de especímenes. Este método se utilizó exitosamente en el área de inspecciones de software (Briand et al. 2000; Thelin 2004; Petterson et al. 2003; Wohlin y Runeson 1998). Aplicar estas técnicas para tener una idea acerca de cuántos requisitos permanecen sin modelar, permitirá reducir la cantidad de requisitos ocultos, mejorando tal vez las heurísticas del proceso de IR.

En trabajos previos (Doorn y Ridao, 2003; 2008; Ridao y Doorn, 2006, 2007a, 2007b), y durante la ejecución del proyecto 03/C248 se han aplicado técnicas cuantitativas para estimar la cantidad máxima de términos que podría contener el LEL y el modelo de Escenarios de un UdeD determinado utilizando una estrategia de captura/recaptura. El análisis de estos modelos, confirmó casi totalmente lo predicho por (Wohlin; 1998) en el caso de inspecciones de software. Cuando el número de observaciones independientes crece, es razonable suponer que la diferencia entre la cantidad estimada y la cantidad efectivamente obtenida se reduce. Se observó también que al aumentar el número de captores aumenta la precisión obtenida al ajustar los datos con una curva, tanto para LEL, como para escenarios. (Hadad et al. 2014)

Transformación ATL de Modelos de Requisitos a Modelos del Negocio en MDD

En otra línea de trabajo, pero aprovechando la información contenida en los modelos de LEL y escenarios, se ha trabajado también en la incorporación de estos modelos a un proceso de desarrollo basado en MDA (Model Driven Architechture). Los métodos de desarrollo basados en este paradigma están comenzando a proveer métodos para la construcción de modelos conceptuales, tomando como entrada modelos de requisitos. En MDA el CIM (Computer Independant Model) es utilizado para representar el modelo del proceso de negocio. Si bien la construcción completamente automática del CIM no es posible, se ha propuesto el uso de los modelos de LEL y escenarios para la derivación de un CIM preliminar. Mediante transformaciones ATL (Atlas Transformation Language), se obtiene un diagrama de clases UML a partir de los modelos de requisitos (Leonardi et al. 2011; 2015). Esta transformación incluye, además, la trazabilidad entre los modelos origen y destino de la transformación. (Felice et al. 2014).

METODOLOGÍA

- Aplicación de teoría de grafos en el modelo del LEL con el fin de detectar agrupamientos de símbolos que permitan mejorar el nivel semántico del modelo.
  - Aplicación de métodos dirigidos por fuerzas a la visualización de los grafos correspondientes al LEL de diferentes casos de estudio.

- Incorporación de técnicas que permitan la visualización del grafo en más de dos dimensiones. Con este fin, como un primer enfoque se propone:
  - Proyectar dinámicamente las n dimensiones del grafo sobre planos arbitrarios.
  - Cambiar automáticamente el tamaño, color y forma de los nodos y arcos, utilizando atributos extraídos del modelo de origen del grafo.
- Determinación de una métrica de calidad de visualización del grafo de un LEL que permita ajustar los parámetros del método dirigido por fuerzas, entre ellos fuerzas aplicadas, constantes de ajuste para dichas fuerzas, ponderación de los vínculos y número de ciclos de iteración.

- Aplicación de otras técnicas de detección de agrupamientos a los léxicos de diferentes casos de estudio. Por ejemplo:
  - Minería de textos aplicada sobre las descripciones textuales de los símbolos del LEL
  - Técnicas de crecimiento de regiones, aplicadas al grafo del LEL
- Aplicación de los algoritmos ya desarrollados, y los resultantes de las mejoras propuestas, a otros modelos del proceso de Ingeniería de Requisitos: Modelo de Escenarios Actuales, Modelo de Escenarios Futuros y Documento de Especificación de Requisitos (SRS)
- Aplicación de métodos de visualización interactivos a los grafos de los diferentes modelos, con el fin de incorporar interacciones que permitan:
  - Efectuar un zoom semántico sobre cada nodo, accediendo al nombre y descripción del elemento correspondiente, manteniendo el grafo en un segundo plano, como contexto.
  - Modificar el aspecto (color, forma, tamaño) de los nodos visualizados según los grupos a los que pertenecen, u otros criterios, como por ejemplo, en el caso del LEL, tipo de símbolo (Sujeto, Objeto, Verbo o Estado).
- Estudio de técnicas que permitan determinar el número de agrupamientos presentes en un modelo, una vez aplicado un algoritmo para detectarlos.
- Diseño de transformaciones para mejorar la heurística de creación de los diferentes modelos del proceso de requisitos, que permitan generar automáticamente una versión, al menos preliminar, de cada modelo a partir del anterior.
- Especificación del ELE incluyendo todos los aspectos del modelo conceptual de datos, proponiendo sólo cambios menores en la notación de las reglas conocidas para la construcción del diagrama correspondiente al modelo, y extendiendo el algoritmo de transformacion para resolver cuestiones relativas a ambigüedades y falta de ortogonalidad.
- Incorporación de cualidades de completitud que se pretenden de la metodología, para posibilitar que el usuario experto pueda optar por una transformación tanto hacia un modelo relacional puro como a uno post-relacional (con características de objetos).
- Definición de las reglas de especificación que establecen la correspondencia ELE con el modelo físico en SQL4.
- Definición de los mecanismos para la reingeniería del ELE, produciendo el mismo modelo conceptual de partida, sin pérdida ni ambigüedades.
- Estudio y formulación de las transformaciones que modifican las restricciones de integridad de datos en un enfoque centralizado, derivadas según un enfoque distribuido mediante segmentación horizontal (sharding) y vertical (proyección controlada). Análisis de la aplicación de los mismos algoritmos mencionados a esquemas particionados, desagregados, replicados, etc. cuyas necesidades de rendimiento requerirán la definición de parámetros de escalabilidad, y reformulación de la consistencia de datos.
- Análisis de la factibilidad de aplicar las capacidades de manejo geométrico de los SIG actuales (en particular open source) a otros tipos de datos no convencionales tales como landmarks morfométricos, extendiendo los tipos de datos ofrecidos por la base de datos, para soportar funciones específicas del dominio.
- Incorporar a SIG existentes módulos de consulta que permitan extraer información derivada para la toma de decisiones y futuras investigaciones.

PLAN DE ACTIVIDADES TOTALES, ESTADO DE AVANCE DE LA LÍNEA Y CRONOGRAMA

El plan de actividades propuesto continúa los desarrollos reportados oportunamente a esa Secretaría, que se encuentran en el siguiente estado de avance:

- Se ha propuesto un algoritmo que permite detectar agrupamientos de símbolos en el LEL y propone métricas para estimar el número de grupos detectados. El análisis de los resultados obtenidos hasta el momento, mediante la aplicación a diversos casos de estudio, indica que es posible efectuar una segunda lectura de algunos de los documentos de IR mediante el estudio de las estructuras semánticas subyacentes (Ridao y Doorn, 2008, 2011, 2013, 2014, en prensa).
- Se han aplicado modelos cuantitativos al estudio de la completitud de modelos de requisitos (Doorn y Ridao 2008)(Hadad, Litvak, Doorn y Ridao 2014).
- Se han diseñado transformaciones ATL para obtener un CIM a partir de los modelos de LEL y Escenarios (Leonardi et al., 2011, 2015). Se ha incorporado un modelo de trazabilidad a la transformación ATL mencionada (Felice et al, 2014).
- Se ha concluido la definición formal de la representación de todos los aspectos incluidos en el modelo conceptual de datos en el ELE, tomando como base el espíritu de la propuesta de Pieris (2013a). Se han propuesto cambios menores en la notación de la mayoría de las reglas contempladas por los algoritmos de conversión convencionales. Sin embargo, y dentro del mismo lineamiento, es posible incorporar con sencillez aspectos sumamente valiosos para la consistencia de los datos, tales como las cardinalidades mínimas y máximas de las relaciones ternarias. Este tipo de relaciones no son consideradas en prácticamente ninguna herramienta de diseño conceptual actual.
- Se ha propuesto un algoritmo formal de transformación que permita obtener un ELE sin pérdida alguna de los conceptos que se plasmaron en el modelo conceptual y que una reingeniería del ELE produzca el mismo DERExt del cual se partió sin pérdida y sin ambigüedades.

- Se ha comenzado a construir una herramienta CASE, original, didáctica y extensible que permite seguir ell ciclo de diseño de una base de datos desde el modelo conceptual, siguiendo los lineamientos de la propuesta que se menciona.
- Se han obtenido de manera automática scripts compilables de SQL4 dada la simplicidad del algoritmo de derivación.
- En lo que respecta a este trabajo, se cree que esta forma de describir un ELE para una base convencional puede ser extendida con relativa sencillez aplicándola a la descripción de la arquitectura completa de los datos en Big Data. En tal sentido, se han probado vía simulación las consecuencias de la segmentación (sharding) de bases de datos, a los fines de encontrar espacios de desarrollo en los que se puedan aprovechar los resultados obtenidos relativos a las DCCV (Dependencias de Comparación de Conjuntos de Valores), dado que una de las principales desventajas de esa práctica es la pérdida de ese tipo de dependencias provocada por la distribución de los datos en diferentes espacios así se han definido formalmente la mayoría de los mecanismos de integridad aplicables a tablas segmentadas.
- Se han desarrollado herramientas que generan automáticamente aplicaciones web y móviles a partir de capas de GIS.
- Una solución para agilizar y mejorar la etapa de entrada de información a un GIS por medio del desarrollo de una herramienta que permita la generación automática de formularios web y dispositivos móviles por medio de los cuales sea posible la carga de datos, en el lugar donde estos se capturan.
- Esta solución permitiría la consistencia de los datos relevados y, aún bajo la existencia de múltiples fuentes de recolección, la integridad de los mismo dado que todos responde al mismo patrón de metadatos.

El plan de actividades previsto incluye las siguientes actividades (tiempo estimado de desarrollo entre paréntesis):

- Mejorar la estrategia de determinación de la presencia de agrupamientos semánticos en un LEL, considerando los mecanismos ya utilizados y proponiendo otros diferentes (18 meses).
- Mejorar los algoritmos previamente utilizados en la visualización de los grafos correspondientes a un LEL en aspectos relacionados con: selección de las fuerzas aplicadas, dimensiones para la construcción del grafo, técnicas de visualización de los grafos resultantes, métricas para evaluar la calidad de la visualización, indicadores del número de agrupamientos presentes en el grafo, ponderación de los vínculos según diferentes criterios, etc. (12 meses que coincidirán total o parcialmente con el punto anterior).
- Aplicar los algoritmos utilizados previamente, y los nuevos mecanismos propuestos, a la visualización de los grafos correspondientes al LEL de nuevos casos de estudio, comparando los resultados obtenidos, con el fin de determinar cuál estrategia resulta más adecuada. (6 meses que coincidirán total o parcialmente con el primer punto).
- Aplicar los algoritmos previos con sus correspondientes mejoras, y los nuevos métodos, a otros modelos del proceso de Ingeniería de Requisitos (6 meses).
- Estudiar posibles mejoras en la estrategia de transformación de modelos de requisitos en un modelo conceptual de datos. (3 meses).
- Proponer transformaciones que permitan automatizar parcialmente las heurísticas de creación de los diferentes modelos del proceso de requisitos (6 meses).
- Incorporar a la notación propuesta los aspectos relativos a cardinalidades máximas y mínimas de los vínculos semánticos unarios, binarios y ternarios. Esto permitirá la definición de patrones de piezas de software destinadas al mantenimiento de restricciones de multiplicidad y sets de dependencias no necesariamente basadas en claves. (4 meses)
- Definición de los aspectos operacionales del mecanismo de reingeniería sin pérdida. (6 meses)
- Incorporación paulatina de estos conceptos en la herramienta CASE hasta el momento implementada parcialmente. A los objetos se agregarán próximamente las relaciones simples y finalmente la complejas a los efectos de contar con una herramienta que en un futuro pueda ser utilizada para la docencia del tema.(12 meses)
- Aplicar la metodología mencionada en un nivel superior de desarrollo, aplicando los conceptos en los casos de BigData. (12 meses)
- Estudiar las características comunes y diferenciales que permitan el tratamiento de datos no convencionales de un SIG, en otros que pueden ser visualizados también según sus características geométricas. (12 meses). Focalizar los siguientes aspectos relacionados a:
    - la naturaleza de los problemas de integridad en SIG, y la posibilidad de generalizar reglas basadas en comparaciones de elementos geométricos.
    - expansión de las capacidades geográficas de los SIG para establecer consultas genéricas que puedan dar lugar a la generación automática de mapas temáticos.
    - Estudiar las generalidades del modelado conceptual de datos morfométricos para poder ser soportados dentro de bases de datos con capacidades espaciales, basicamente GIS.

APORTES ACADÉMICOS Y DE TRANSFERENCIA ESPERADOS

Los conocimientos adquiridos se vuelcan en el dictado y confección de material didáctico para el área de Bases de Datos y Estructuras de Almacenamiento, de la carrera Ingeniería de Sistemas. Se focaliza especialmente en la asignación de proyectos para la aprobación del examen final y proyectos de cátedra que incluyan problemáticas actuales; también en materias optativas (Ingeniería de Requisitos), y el curso de posgrado Tópicos de Ingeniería de Requisitos de la Maestría en Informática Avanzada de la UNLaM, así como para: Sistemas de Información Geográfica Open Source, materia optativa de grado en las carreras de Ingeniería de Sistemas y Licenciatura en Matemáticas de la Fac de Cs. Exactas UNCPBA, y las materias Estructuras de Datos e Introducción a las Bases de Datos de la Tecnicatura Universitaria en Programación y Administración de Redes (TUPAR) y de la reciente Tecnicatura Universitaria en Desarrollo de Aplicaciones Informáticas. El material es referencia bibliográfica para 'Tópicos de Ingeniería de Requerimientos' en la Maestría en Informática avanzada de la Fac. de Informática de la UNLaM.

Con temáticas surgidas de estas líneas de trabajo, varios grupos de alumnos se encuentran desarrollando el proyecto final de carrera, con diferente grado de avance en su trabajo.

Se está trabajando en colaboración con el Grupo de Evolución Morfológica de la Facultad de Ciencias Naturales y Museo de la Universidad Nacional de La Plata, en el desarrollo de aplicaciones de software para diferentes análisis morfométricos y se han realizado presentaciones conjuntas en congresos y revistas del área.

En el marco de esta línea, los aspectos relacionados con la aplicación de técnicas cuantitativas al proceso de requisitos forman parte del plan de trabajo de Marcela Ridao para la tesis del doctorado en Ciencias Informáticas de la Universidad Nacional de La Plata, cuyo título es: Técnicas Cuantitativas Orientadas al Reuso Semántico de Modelos de Requisitos.

ANTECEDENTES DEL GRUPO EN LA TEMÁTICA

La presente línea continúa la correspondiente del proyecto Bases de Datos y Procesamiento de Señales 03/C248, que finaliza en diciembre del corriente año y la interdisciplinariedad permite ver la realidad desde distintas aristas, de ahí su importancia en la universidad como práctica fundamental para avanzar en diferentes áreas.

Se han publicado artículos en revistas y actas de congresos nacionales e internacionales, mencionados en las secciones anteriores, e incluidos en la sección de Referencias y Principal Bibliografía.

REFERENCIAS Y PRINCIPAL BIBLIOGRAFÍA SOBRE LA LÍNEA

- Aiello, A., Silveira, A. (2004). Trazado de grafos mediante métodos dirigidos por fuerzas: revisión del estado del arte. Tesis de Licenciatura en Ciencias de la Computación.
- Badía, A. (2004). Entity-Relationship Modeling Revisited, SIGMOD Record, vol. 33 no.1. pp 77-82, 2004.
- Badia, A. and Lemire, D. (2011). A call to arms: revisiting database design. SIGMOD Record 40 (3), pages 61-69.
- Berry, D.M., Kamsties, E. (2004). Ambiguity in Requirements Specification, en Perspectives of Software Requirements. Kluwer Academic Publishers. Norwell, Massachusetts. USA. ISBN 1-4020-7625-8. pp 7-44.
- Briand, L. El Emam, K., Freimut, B., Laitenberger, O. (2000). A Comprehensive Evaluation of Capture-Recapture Models for Estimating software Defects Contents. IEEE Transactions on Software Engineering, Vol. 26, N° 6. pp 518-540. doi: 10.1109/32.852741.
- Camps, R. (2002). Transforming n-ary relationships to database schemas: an old and forgotten problem. Research Report LSI-02-5-R. Univ. Politécnica de Catalunya, España.
- Chen, P.P.(1990). The Entity-relationship Model: Toward a Unified View of Data, ACM Trans. on Database Systems, vol. 1, no. 1, pp 9-36, Codd, E., The Relational Model for Database Management. Version 2, Addison Wesley Publ. Co.
- CodeFuture Corporation. (2008). Cost-effective Database Scalability using Database Sharding. Disponible en: http://codefutures.com/database-sharding/
- Cuadra, D., Nieto, C., Martínez, P., Castro, E., Velasco, M. (2002). Preserving Relationship Cardinality Constraints in Relational Schemata, In Database Integrity: Challenges and Solutions, Idea Group Inc., Jorge Doorn and Laura Rivero Editors, pp. 66-112.
- Das, R., Kalita, J., Bhattacharyya, D.K. (2011). A pattern matching approach for clustering gene expression data, Int. J. Data Mining, Modelling and Management, Vol. 3, No. 2, pp 130-149.
- Date, C. (2000). An Introduction to Database Systems, Addison Wesley.
- dbShards (2015). dbShards for Scaling Big Data with Database Sharding. Disponible en: www.dbshards.com/
- Doorn, J., Ridao, M. (2003). Completitud de glosarios: un estudio experimental. Proceedings de WER03 - Workshop em Engenharia de Requisitos, Piracicaba-SP, Brasil, Noviembre 27-28, pp 317-328.
- Doorn, J., Ridao, M. (2008). Completeness Concerns in Requirements Engineering. En KHOSROW-POUR M. (editor) Encyclopedia of Information Science and Technology, 2nd edition. IGI Global, Hershey (USA). pp. 619-624. ISBN: 978-11-60566-026-4.
- Duan, C., Laurent, P., Cleland-Huang, J., y Kwiatkowski, C., (2009). Towards automated requirements prioritization and triage, Requirements Engineering Journal, 14, 2, pp.73-89.
- Eades, P. (1984). A heuristic for graph drawing. Congressus Numerantium 42, 149–160
- Felice, L., Ridao, M., Leonardi, M.C., Mauco, M.V. (2014). Enhancing an ATL Transformation with Traceability. En Proceedings de SEAS 2014 – 3rd. International Conference on Software Engineering and Applications.
- Goelman, D., Hussmann, H.(1999). Using UML/OCL Constraints for Relational Database Design, In UML'99: The Unified Modeling Language - Beyond the Standard, Lecture Notes in Computer Science, Springer, 1723, Robert B. France and Bernhard Rumpe Editors, pp. 598-613.
- Goelman, D., Song, I-Y. (2004). Entity-Relationship Modeling Re-revisited, ER 2004, Proceedings 23rd International Conference on Conceptual Modeling, Shanghai, China. Lecture Notes in Computer Science, Springer, 3288, Paolo Atzeni, Wesley W. Chu, Hongjun Lu, Shuigeng Zhou, Tok Wang Ling Editors, pp. 43-54.
- Hadad, G. (2007). Uso de Escenarios en la Derivación de Software. Tesis doctoral. Facultad de Cs Exactas, Univ. Nacional de La Plata.
- Hadad, G., Doorn, J., Kaplan, G. (2009). Explicitar Requisitos del Software usando Escenarios. Proceedings de WER'09 – XII Workshop on Requirements Engineering. Valparaíso, Chile, Julio 16-17, pp 63-74.
- Hadad, G., LItvak, C., Doorn, J., Ridao, M. (2014). Dealing with Completeness in Requirements Engineering. En En KHOSROW-POUR M. (editor) Encyclopedia of Information Science and Technology, 3rd edition. IGI Global, Hershey (USA). pp. 2854-2863. ISBN: 9781466658882.
- Halpin, T.(2001). Information Modeling and Relational Databases. From Conceptual Analysis to Logical Design", Morgan Kaufmann Publishers.
- Henley, S. (2006).The problem of Missing Data in Geosciences databases. Computers and Geosciences. Elsevier.
- Leite, J.C.S.P., Doorn, J.H., Hadad, G.D.S., Kaplan, G.N., Ridao, M. (2004). Defining System Context Using Scenarios. Perspectives of Software Requirements. Kluwer Academic Publishers. Norwell, Massachusetts. USA. ISBN 1-4020-7625-8.
- Leite, J.C.S.P., Hadad, G.D.S., Doorn, J.H., Kaplan, G.N. (2000). A Scenario Construction Process. Requirements Engineering Journal, Vol.5, N° 1, pp 38-61. doi: 10.1007/PL00010342.
- Leite, J.C.S.P., Rossi, G., Balaguer, F., Maiorana, V., Kaplan, G., Hadad, G., Oliveros, A. (1997). Enhancing a Requirements Baseline with Scenarios. Requirements Engineering Journal, Vol.2, N° 4, pp 184-198. doi: 10.1109/ISRE.1997.566841.
- Leonardi, M.C., Ridao, M., Mauco, M.V., Felice, L., Montejano, G., Riesco, D., Debnath, N. (2011). An ATL Transformation from Natural Language Requirements Models to Business Models of a MDA Project. Proceedings ITST 2011. Saint-Petersburg, Russia.

. Leonardi, M.C., Ridao, M., Mauco, M.V., Felice, L. (2015). A Natural Language Requirements Engineering Approach for MDA. International Journal of Computer Science Engineering and Applications IJCSEA 5(1).. AIRCC Publising Corporation. Pp 1-18. ISSN: 2230-9016 (Online) 2231-0088 (Print)
. Mo, Y., Cao, Z., Wang, B. (2012). Occurrence-Based Fingerprint Clustering for Fast Pattern-Matching Location Determination. Communications Letters, IEEE 16(12), pp 2012-2015.
. Otis, D.L., Burnham, K.P. White G.C., Anderson D.R. (1978). Statistical inference from Capture on Closed Animal Populations, Wildlife Monograph, 62.
. Petersson, H., Thelin, T., Runeson, P., Wohlin, C. (2003). Capture–recapture in software inspections after 10 years research—theory, evaluation and application. The Journal of Systems and Software, 72, pp 249–264. doi: 10.1016/S0164-1212(03)00090-6.
. Pieris, D. (2013a). Modifying the entity relationship modeling notation: towards high quality relational databases from better notated ER models. arXiv preprint arXiv:1306.5690.
. Pieris, D. (2013b). A novel ER model to relational model transformation algorithm for semantically clear high quality database design. arXiv preprint arXiv:1306.6734.
. Pieris, D., & Rajapakse, J. (2012). Logical database design with ontologically clear entity relationship models. Paper presented at the Information And Automation For Sustainability Beijing, China.
. Ridao, M., Doorn, J. (2006). Estimación de Completitud en Modelos de Requisitos Basados en Lenguaje Natural. Proceedings de WER06 - Workshop em Engenharia de Requisitos, Rio de Janeiro, Brasil, pp 151-158.
. Ridao, M., Doorn, J. (2007a). Hipótesis Bayesiana en Modelos de Completitud. Proceedings de WICC 2007 – IX Workshop de Investigadores en Ciencias de la Computación. Trelew, Argentina, Mayo 3-4, pp 380-384.
. Ridao, M., Doorn, J., Cabuya, M., Kaplan, G. (2007b). Simulación de Modelos de Completitud en Ingeniería de Requisitos. Anuario de Investigaciones, Resúmenes Extendidos. Universidad Nacional de la Matanza, pp 61-65.
. Ridao, M., Doorn, J. (2008). Mejorando el Nivel Semántico del Léxico Extendido del Lenguaje. Proceedings de WICC 2008 – X Workshop de Investigadores en Ciencias de la Computación. General Pico, Argentina, Mayo 5-6, pp 419-423.
. Ridao, M., Doorn, J. (2011). Agrupamientos Semánticos en Glosarios del Universo de Discurso. Proceedings de WICC 2011 – XIII Workshop de Investigadores en Ciencias de la Computación., Rosario, Argentina – Mayo 5-6, pp. 268-272. ISBN: 978-950-673-892-1.
. Ridao, M., Doorn, J. (2013). Semántica Oculta en Modelos de Requisitos. En: XV Workshop de Investigadores en Ciencias de la Computación, WICC. Paraná, Argentina.
. Ridao, M., Doorn, J. (2014). Detección de Agrupamientos en Glosarios del Universo de Discurso. En CONAIISI 2014, Segundo Congreso Nacional de Ingeniería Informática. San Luis, Argentina.
. Ridao, M., Doorn, J. (en prensa). Agrupamientos en Glosarios del Universo de Discurso. Revista Tecnología y Ciencia. Universidad Tecnológica Nacional. ISSN: 1666-6917.
. Rivero, L. (2008). Equality Dependencies: Analysis, Modeling and Conceptual Issues". International Journal of Computer Science and Software Technology - IJCSST, International Science Press, Nueva Delhi, India, tomo 1, no 1, pp. 15–25. ISSN 0974 3898.
. Rivero, L. (2009). Referential Constraints. En KHOSROW-POUR M. (editor) Encyclopedia of Information Science and Technology, 2nd edition. IGI Global, Hershey (USA). Vol VII. pp. 3251-3257. ISBN: 978 11 60566 026 4.
. Rivero, L. (2011). Respecification of Atypical Referential Constraints. Aceptado para publicación en International Journal of Database Management Systems (IJDMS). Academy & Industry Research Collaboration Center (AIRCC).
. Rivero, L., Doorn, J. and Ferraggine, V. (2004). Enhancing Relational Schemas Through the Analysis of Inclusion Dependencies. International Journal of Computer Research, 12(4), Nova Publishers.
. Rumbaugh, J., Jacobson, I., Booch, G. (1999). The Unified Modeling Language Reference Manual, Addison Wesley, 1999.
. Teorey, T.J. (1990). Database Modeling and Design. The Entity-Relationship Approach, Morgan Kaufmann Publishers, 1990.
. Teorey, T., Lightstone, S., Nadeau, T., Jagadish, H.V. (2011). Database Modeling and Design. Logical design. 5th edition. Morgan Kaufmann Publishers.
. Thalheim, B. (2000). Entity-Relationship Modeling. Foundations of Database Technology. Springer-Verlag.
. Torcida, S., Perez, S.I. & Gonzalez, P. N. (2014). An Integrated Approach for Landmark-Based Resistant Shape Analysis in 3D. Evolutionary Biology, 41(2), 351-366.
. Walshaw, C. (2003). A multilevel algorithm for force-directed graph-drawing. Journal of Graph Algorithms and Applications 7(3), 253-285
. Wohlin, C., Runeson, P. (1998). Defect contents estimation from review data, en Proceedings of the 20th International Conference on Software Engineering, pp 400-409. doi: 10.1109/ICSE.1998.671393.
. Zimmermann, M., Ntoutsi, I., Siddiqui, Z., Spiliopoulou, M., Kriegel, H.P. (2012). Discovering Global and Local Bursts in a Stream of News. In: 27th Annual ACM Symposium on Applied Computing. SAC '12, pp. 807-812. Italy.

Línea 2: "Análisis y Procesamiento de Señales"

Director de línea: WAINSCHENKER, Rubén Sergio
Personal Participante: MASSA, José María
TRISTAN, Paula Mariela
MARONE, José Antonio
CURTI, Hugo.
ZUBELDÍA, Alfonso Román (Colaborador)
MENCHON, Martín (Colaborador)

OBJETIVOS Y/O FINALIDAD

Desarrollar diferentes algoritmos de obtención de información a partir de datos presentes en señales digitales unidimensionales, bidimensionales, o multidimensionales en general, ya sea tanto en tiempo real como diferido, profundizando especialmente el análisis y desarrollo de diversas alternativas de mejora u optimización en las diferentes etapas o aspectos que lo componen.

Objetivos específicos

- Continuar con el estudio de imágenes multidimensionales como las satelitales, avanzando en los procesos de extracción de información espacial a partir del análisis y la posible agrupación de datos similares para su integración directa en Sistemas de Información Geográfica.
- Continuar avanzando en la creación y desarrollo de alternativas de procesamiento de imágenes de diferentes orígenes, que permitan asistir y colaborar en el desarrollo de buenas prácticas en actividades relacionadas a la agricultura de precisión y al proceso agroindustrial.
- Desarrollar técnicas de agrupamiento y clasificación por similitud de datos multidimensionales, de manera autónoma. Así mismo, analizar la elección de diferentes métricas como estimador de semejanza de los métodos de clasificación.
- Continuar con la optimización del cálculo de dosis en radioterapia a través de la reducción del tiempo de cálculo de interacciones, la implementación de distintas técnicas de paralelización e implementación de algoritmos eficientes que aprovechen la arquitecturas subyacente de punto fijo y flotante.
- Segmentación de estructuras de interés en imágenes médicas por medio de técnicas basadas en Machine Learning, particularmente de aprendizaje profundo (Deep Learning) sobre imágenes y videos de ultrasonido intravascular y otro tipo de modalidades.
- Promover el desarrollo y la automatización en la construcción de sistemas de información geográfica con herramientas de código abierto.

ESTADO ACTUAL DEL CONOCIMIENTO

En toda señal puede haber información, en algunos casos es evidente y en otros no lo es; obtener información de las señales es una de las tareas más importantes dentro del Procesamiento de Señales. Habiendo desarrollado técnicas de registración automática en imágenes satelitales, método que puede extenderse a diferente tipo de imágenes [Tristan et al 2012] se continua el estudio de organizar los datos en imágenes satelitales tratando de agrupar los que tienen semejanza entre si en todas las bandas asociadas a datos de la superficie terrestre. Los métodos de agrupación automática no asistida (Clusterización) se pueden aplicar tanto a imágenes satelitales como a bases de datos asociados a información médica, seguridad, planificación urbana agricultura, geográfica, forestación, etc.

En las últimas décadas, se viene adquiriendo cada vez más y más cantidad de datos multidimensionales de distinto tipo. Estos datos provienen de distintos campos del conocimiento, ya sea información genética, imágenes médicas, imágenes geográficas (satelitales), etc. Obtener información a partir de estos datos se realiza con técnicas de Data Mining. Una de las principales técnicas no supervisadas es el clustering. Aunque no se haya acordado en una única definición de cluster, el proceso de agrupar datos parecidos en categorías o clases, se viene realizando con técnicas de partición, de procesos jerárquicos y de estudio de densidades [(Jain & Dubes 1988),(Everitt et al 2011),(Hartigan 1975)]

Los métodos de agrupamiento no supervisado (Clustering) se aplican en muy diversas áreas incluyendo medicina bioinformática, mercados financieros, segmentación de imágenes, busqueda en la WEB, educación, por mencionar algunas de ellas [(Anderberg 1973), (Jiang et al 2004), (Jain et al1999)].

Diferentes definiciones de clustering han dado origen a diferentes algoritmos de agrupamento según el campo en donde se lo ha desarrollado como ser las estadísticas, pattern recognition y machine learning [(Hubert et al 1996), (Fukunaga 1990), (Kodratoff and Michalski 1990)].

El desarrollo de herramientas de software basados en imágenes que asistan a diversas áreas productivas involucran por lo general varios paradigmas: manejo de algoritmos de procesamiento de imágenes para extraer información de interés y gestión de bases de datos con capacidades geográficas y SIG. Con el afianzamiento de estas áreas de investigación (Tristan et al 2009), surge la posibilidad de desarrollar un conjunto de herramientas orientado a la web o sobre plataformas móviles, que permitieran la implementación y el desarrollo de sistemas que permitan, a bajo costo, resolver problemas de aplicación concretos. Particularmente, el sector agropecuario, se ha convertido en una actividad económica cada vez más competitiva, con demandas cada vez más exigentes, precios cada vez más ajustados, exigencia de alimentos de mayor calidad de manera sostenible y respetando el medio ambiente, de manera tal que ha surgido la necesidad de aplicar nuevas tecnologías. En este sentido se ha

avanzado en un proceso agropecuario que permita efectuar una intervención correcta en el momento adecuado y en el lugar preciso.

La determinación de la dosis de radiación absorbida en tejido humano es de suma importancia para lograr un tratamiento de radioterapia eficaz. El método más apropiado para estimar la dosis absorbida es el cálculo basado en Monte Carlo, considerado como el más preciso. Los programas que realizan este cálculo utilizando diferentes estrategias han debido optar por alguna de las variantes en el compromiso, aún no resuelto apropiadamente, entre la calidad de la estimación y el costo temporal del cálculo involucrado. Muchas de las técnicas existentes de reducción de tiempo se basan en una simplificación del problema y acarrean una pérdida de calidad en los resultados. Otras técnicas, entre las que se encuentran el cálculo directo, el precálculo y la paralelización permiten reducir el tiempo de cálculo sin perder calidad en los resultados realizando un cálculo completo sin simplificaciones. Entre estas últimas se pueden mencionar trabajos realizados con la técnica de Grid Computing y el uso de GPU (de Greef, et al., 2010; Jahnke, et al., 2012; Xun, 2010).

En trabajos anteriores (Massa et al., 2011) se ha optimizado el tiempo de cálculo reemplazando la función de rechazo por métodos de cálculo directo. Considerando interacciones de fotones con el medio se han logrado factores de reducción que van entre 7 y 30, manteniendo la calidad de los resultados respecto de los métodos tomados como referencia. Además, se ha aplicado una estrategia de paralelización altamente inmune a fallas (Massa, et al. 2010), con la que se obtuvo la cantidad óptima de unidades de cálculo que reduce el tiempo global de cálculo. Esto se realizó mediante una técnica ad-hoc y mediante el uso de la técnica de Grid Computing.

Para acelerar el cálculo de dosis por el método de Monte Carlo, además de las soluciones basadas en arquitecturas de tipo X86, en los últimos años se han implementado soluciones para otras arquitecturas como las de las Unidades de Procesamiento Gráfico (GPU) y Field Programmable Gate Array (FPGA) (Li, et al. 2013), (Fanti, et al. 2009), (Luu, et al. 2009). Las soluciones implementadas sobre FPGA aprovechan la velocidad y el paralelismo de este tipo de hardware para realizar cálculos de punto fijo, pero se deben implementar con cuidado de modo de reducir la pérdida de precisión de los resultados en comparación a las soluciones de punto flotante implementadas sobre arquitecturas de tipo X86.

Respecto a la aplicación de nuevos métodos de muestreo que reduzcan la cantidad de rechazos, además de los métodos directos es posible aplicar métodos de reducción de rechazos como Ziggurat (Buchmann, et al. 2014). Si bien estos métodos son más eficientes que sus contrapartes con rechazos, aún dependen de una implementación que aproveche los recursos de la arquitectura subyacente.

La digitalización de las diferentes modalidades de imágenes médicas y su disponibilidad en sistemas de almacenamiento, junto con la adopción de estándares para comunicación, facilita el desarrollo de herramientas de procesamiento digital de imágenes con el fin de brindar soluciones de asistencia al diagnóstico (Perino, et al. 2012).

Respecto de la aplicación de métodos de Machine Learning, particularmente se han utilizado redes neuronales convolucionales (CNN) que son modelos matemáticos que utilizan una colección de elementos computacionales simples llamados neuronas que observan una pequeña porción de una imagen. Estas redes se inspiran en los procesos biológicos que ocurren en el cerebro imitando en cierta manera el funcionamiento de la corteza visual humana (Schmidhuber, et al. 2015). Estas redes han revolucionado el mundo de la visión por computador como puede comprobarse en (Agrawal, 2014) en gran medida gracias a una nueva concepción arquitectónica de las mismas y al aumento en la potencia de cálculo brindada por las placas gráficas (GPU). Las imágenes médicas han sido un dominio poco explorado en la aplicación de algoritmos basados en Deep Learning (Bengio, et al. 2013). Esto tal vez se deba a que existen varios desafíos, tales como la existencia de diferentes modalidades de imágenes, la falta de información etiquetada por expertos (Ground Truth) y los defectos propios de las imágenes médicas.

METODOLOGÍA

En Teledetección, los mayores esfuerzos de investigación se han enfocado en mejorar la calidad de la información contenida en las imágenes obtenidas a través de sensores remotos. En ese sentido los métodos de fusión de datos, parten de imágenes de diferente resolución espacial, espectral y/o radiométrica y obtienen imágenes mejoradas. Otra proceso en imágenes satelitales, que es común con el tratamiento de datos multivariados, es la clusterización, o sea la clasificación no supervisada. Se ha actualizado la bibliografía de clasificación y clusterización y se han implementado los métodos más utilizados aplicándolos a imágenes satelitales y otros datos multidimensionales. Se ha hecho un estudio de los criterios de similitud empleados, particularmente la medición de cercanía de puntos en un espacio multidimensional. Se considera que los métodos actuales para clusterizar pueden ser mejorados sustancialmente, en tal sentido se ha comenzado con la elección de una métrica adecuada al problema a tratar utilizando los métodos clásicos de clusterización como k-means y se ha comenzado a realizar pruebas buscando automatizar la elección de los parámetros para estudiar densidades en el espacio de las características.

Otro aspecto interesante de análisis, es la determinación de la calidad y clasificación de la cosecha. La calidad de los granos define las condiciones de la comercialización, por ende el valor de comercialización y un posible incremento en el ingreso de divisas al país. Actualmente la determinación de la calidad de los granos es realizada en forma manual por los Peritos Clasificadores de Granos, los cuales se basan en Normas y Estándares Oficiales. La posibilidad de contar con un software que permita realizar esta tarea en forma automática, a partir de una imagen de la muestra, con mucha mayor precisión que el "buen ojo" de los peritos, sería de gran ayuda a la tarea de clasificación. Además, los avances tecnológicos, posibilitan, por ejemplo, contar con un software de clasificación en un dispositivo móvil, lo que permitiría que un análisis preliminar de la calidad del grano se haga en el lugar mismo de la cosecha, pudiendo eventualmente cambiar la configuración de la máquina cosechadora con el objetivo de mejorar la calidad del material obtenido. Siguiendo la cadena de comercialización de granos, se ha planteado la necesidad de poder detectar o clasificar semillas de diferentes granos de manera automática cuando éstas están siendo transportadas para su almacenamiento. En las grandes plantas de acopio de cereales, las semillas son transportadas entre silos mediante cintas transportadoras que se desplazan rápidamente, cualquier error humano que implique una mezcla de granos

resultaría en una importante pérdida de dinero. Una detección en tiempo real de la mercadería transportada y de la no presencia de semillas de diferente variedad resultaría de gran importancia a la hora del movimiento y almacenamiento de cereales.

Respecto al cálculo de dosis en radioterapia, los beneficios de implementar soluciones de punto fijo sobre las de punto flotante radican en que las primeras utilizan menor cantidad de área y recursos que las últimas. Esto impacta sobre la latencia general y el área de diseño. Las técnicas de optimización de datos se convierte en algo fundamental para optimizar tiempo y área. Sin embargo hay que tener en cuenta que el uso de punto fijo y la reducción la cantidad de bits empleada para la representación impactan en la precisión, introduciendo errores de cuantificación (Shi, et. 2013)(Boland, et. al 2013). Para estimar la cantidad adecuada de bits necesarios para la representación en punto fijo, se requiere realizar un análisis de la precisión mínima tolerada por el problema y su rango de valores. Este análisis se puede realizar mediante análisis matemático de las funciones y también mediante simulaciones en donde se comparen dos conjuntos de resultados generados, uno con el algoritmo original de punto flotante y el otro con el algoritmo implementado sobre punto fijo.

Específicamente para el problema del cálculo de dosis por el método de monte carlo, se tomará el algoritmo de punto flotante para la generación de las trayectorias iniciales de las partículas que son emitidas por la fuente de radiación, se realizará un análisis para determinar la cantidad de bits requeridos sobre una arquitectura de punto fijo para tolerar un error del 1% sobre la matriz de cuantificación de dosis. Este análisis se realizará generando un archivo de espacio de fase con el programa PENELOPE, ampliamente utilizado como referencia en el área de aplicación. Luego se planifica implementar un algoritmo de punto fijo sobre FPGA y se procederá a generar un segundo archivo de espacio de fase para punto fijo. Finalmente, una vez estimada la cantidad de bits necesarios para mantener el error por debajo de la cota mencionada, se realizarán estimaciones deaceleración de tiempos de cálculo para las arquitecturas de punto fijo.

En cuanto al método de muestreo basado en Ziggurat se planifica realizar diferentes implementaciones y analizar la utilización de recursos de hardware para cada una de ellas, con el fin de seleccionar la implementación mas eficiente para una arquitectura dada.

En cuanto a la aplicación de métodos de Machine Learning sobre imágenes médicas, se ha comenzado con desarrollo de nuevas arquitecturas de redes neuronales para la segmentación automática de borde de lúmen en imágenes de IVUS. Se planifica seguir explorando arquitecturas de redes neuronales con el objetivo de hallar una arquitectura específica para la detección y segmentación de imágenes IVUS. Además se planifica extender su aplicación a otras modalidades de imágenes como por ejemplo imágenes dermatológicas.

PLAN DE ACTIVIDADES TOTALES, ESTADO DE AVANCE DE LA LÍNEA

El plan de actividades propuesto continúa los desarrollos iniciados en la Línea 2 del Proyecto de Incentivos que está culminando, cuyos resultados fueron:

- Procesamiento de imágenes satelitales:
  - Se ha profundizado el estudio de técnicas de clasificación en general y de clasificación no supervizada en particular tanto en imágenes satelitales, como en bases de datos multivariadas.Se han desarrollado técnicas de clasificación automática de cubiertas de la superficie terrestre con imágenes satelitales como en datos multidimensionales.
  - Se han hecho estudios de la incidencia de la elección de diferentes métricas como criterio de similitud en la clasificación automática de datos multidimensionales.
- Cálculo de Dosis en Radioterapia:
  - Se ha comenzado el diseño de una plataforma web para el acceso al sistema de cálculo en paralelo basado en Grid Computing.
  - Se ha finalizado la escritura de una publicación respecto a la optimización del fenómeno de producción de pares.
  - Se ha comenzado a evaluar la posibilidad de implementar técnicas que utilicen arquitecturas de punto fijo.
  - Se ha analizado desde el punto de vista de arquitecturas de hardware, la performance de algoritmos de muestreo como el método de Ziggurat.
- Se ha construido una arquitectura de red neuronal para la segmentación de borde de lúmen en imágenes IVUS.
- Se ha colaborado en la implementación de herramientas para su aplicación en problemas de salud y medioambiente.

Se planea desarrollar las siguientes actividades:

- *Procesamiento de imágenes*
  - Avanzar en los procesos de extracción automática de información de la superficie terrestre.
  - Continuar con el estudio de metodologías automáticas de clustering y elección adecuada de la métrica a utilizar como criterio de similitud o de proximidad en espacios multidimensionales.
  - Avanzar en el desarrollo de aplicaciones orientadas al proceso Agroindustrial
- *Sistemas de Información Geográfica*
  - Continuar con la divulgación y capacitación en el desarrollo de SIG utilizando herramientas de Open Source.
  - Continuar avanzando en el desarrollo de nuevas capacidades de los SIG, extendiendo por ejemplo métodos de clasificación estándares a aquellos que incluyan características de los objetos y su entorno geográfico.
  - Avanzar en la automatización de la construcción de capas de información geográfica provenientes de la información recogida por sensores de levantamiento del relieve de la superficie del lecho submarino.
- *Cálculo de Dosis en Radioterapia*
  - Implementar una solución en una arquitectura de punto fijo sobre un hardware tipo FPGA Xilinx Vivado HLS, determinando la cantidad óptima de bits necesarios para la representación numérica mediante punto fijo, teniendo en cuenta el rango y la precisión.
  - Medición de la performance lograda y el factor de optimización de las arquitecturas de punto fijo respecto a las de punto flotante.

- Realizar un relevamiento del desempeño del método de muestreo de Ziggurat sobre diferentes arquitecturas de hardware.
- Extender la red neuronal construida para detectar borde de lúmen en imágenes IVUS, para la detección de otras estructuras de interés y continuar con el relevamiento bibliográfico del tema.

APORTES ACADÉMICOS Y DE TRANSFERENCIA ESPERADOS

El plan de trabajo presentado es de interés nacional por la inmediata usabilidad de sus resultados en el medio informático y social de nuestro país. En cuanto a la formación de recursos humanos se trabaja permanentemente en el desarrollo de tesinas de grado como así también se están desarrollando tesis de posgrado (Maestrías y Doctorados).

Los aportes de investigación se vuelcan en los cursos "Procesamiento de Sonido", "Procesamiento de Imágenes I", "Procesamiento de Imágenes II", "Sistemas de Información Geográfica Open Source" de la carrera de Ingeniería de Sistemas (como materias optativas en las carreras de Ingeniería de Sistemas y Licenciatura en Matemática) y "*Pattern Recognition*", "Procesamiento de Imágenes Satelitales" y "Teledetección: Fundamentos y su aplicación en sistemas de información geográfica" (como cursos de Postgrado en el Doctorado de Matemática Computacional e Industrial)

Se está trabajando en colaboración con el Grupo de Ecosistemas de la Facultad de Cs. Veterinarias de la UNICEN en el desarrollo de herramientas para la informatización y análisis de las observaciones medioambientales, específicamente en las campañas de relevamiento de información de anátidos. Los trabajos en conjunto se han presentado en numerosas reuniones de las que han participado autoridades y organismos ambientales de la nación. (Álvarez, et al. 2012).

Se está trabajando en colaboración con eL Instituto Nacional de Investigación y Desarrollo Pesquero -inidep- y la Facultad de Ingenierìa en la automatización del proceso de generación de capas de información de un GIS basados en los datos obtenidos por sensores del lecho submarino.

Se está trabajando en colaboración con el Instituto Nacional de Tecnología Agropecuaria INTA, en el desarrollo de aplicaciones de software orientadas a promover el desarrollo agropecuario, atendiendo de esta forma los crecientes requerimientos del sector.

Se espera contribuir a la reducción del tiempo de cálculo de la dosis en radioterapia, que forma parte de una de las líneas de investigación del proyecto. Académicamente se espera incorporar estos avances en la tesis del Doctorado en Matemática Computacional e Industrial del Ing. Luis Pantaleone, bajo la dirección del Dr. José M. Massa y comenzar a escribir una segundo trabajo para enviar a su publicación.

Desde febrero de 2015 se esta trabajando en vinculación con el BCNPCL (Barcelona Perception Computing Laboratory) y por su intermedio con el CVC (Centro de Visión por Computador) para el desarrollo de nuevas arquitecturas de redes neuronales para la segmentación automática de patrones de interés en imágenes médicas. Todos los aportes resultantes de esta vinculación serán incorporados en la tesis doctoral del Ing. José Marone , bajo la dirección del Dr. José M. Massa.

Por otro lado en 2014 se firmo un convenio marco con la corporación INTRAWAY, y se espera firmar un convenio específico durante 2016 para el desarrollo de algoritmos de tracking visual de personas orientado a la seguridad mediante cámaras IP.

Se está trabajando en colaboración con el Hospital Italiano de la Ciudad de Buenos Aires para la implementación de técnicas de detección y seguimiento de mirada (gaze tracking) en sistemas de rehabilitación cognitiva.

ANTECEDENTES DEL GRUPO EN LA TEMÁTICA

El trabajo citado en la presente línea tomará como base los desarrollos obtenidos en la línea 2 del proyecto Bases de Datos y Procesamiento de Señales que finaliza en diciembre del corriente año.

Se han publicado artículos en revistas y actas de congresos nacionales e internacionales (Alvarez et al 2011).

El Laboratorio de Sistemas Digitales Tandil, perteneciente al INTIA y en especial el Ing. Luis Pantaleone posee una amplia experiencia en la implementación de algoritmos eficientes sobre arquitecturas de punto fijo y en especial sobre FPGA. (Vazquez, et al. 2014). Recientemente surgió la posibilidad de aplicar ese conocimiento en la optimización del cálculo de dosis por el método de Monte Carlo.

Se han llevado a cabo diferentes convenios de colaboración con el grupo Arbeitsgruppe für StrahlungPhysik de la Universidad Técnica de Dresden para la colaboración en el trabajo respecto de la reducción de tiempo de cálculo de dosis de radioterapia.

Durante el año 2015, un integrante del Instituto ha realizado una estadía corta de doctorado en el instututo BCNPCL (Barcelona Perception Computing Laboratory) con una beca obtenida mediante el programa BEC.AR para especializarce en el área de Machine Learning, especialmente enfocado a técnicas de aprendizaje profundo (DeepLearning) dicha estadia culminó con la redacción de un artículo científico y la propuesta de una beca de postgrado (CONICET) en el tema, para un alumno avanzado de la carrera de ingeniaria de sistemas y en dirección conjunta con investigadores de dicho instituto.

REFERENCIAS Y PRINCIPAL BIBLIOGRAFÍA SOBRE LA LÍNEA

- (Alvarez et al 2011) M. Álvarez, P. Tristán, J. Massa and R. Wainschenker. Clasificación Automática de Cubiertas Terrestres en Imágenes Satelitales.XVII Congreso Argentino de Ciencias de la Computación..
- (Anderberg-1973) M. R. Anderberg. "Cluster Analysis for Applications". New York: Academic Press, 1973.
- (Everitt et al 2011) B. S. Everitt, S. Landau, M. Leese, and D. Stahl. "Cluster Analysis".West Sussex, UK: Wiley, 2011.
- (Fukunaga 1990) K. Fukunaga. "Introduction to Statistical Pattern Recognition". Boston, MA: Academic Press, 2nd edition.
- (Hartigan 1975) J. A. Hartigan. Clustering Algorithms. New York: Wiley & Sons, 1975.
- (Hubert et al 1996) L. Hubert & P. Arabie. "An overview of combinatorial data analysis". World Scientific Publication, 1996.
- (Jain & Dubes 1988) A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice Hall, 1988.
- (Jain et al1999) A. K. Jain, M. N. Murty, and P. J. Flynn. "data clustering: A review". ACM Computing Surveys, 31(3):264–323, 1999. doi: 10.1145/331499.331504.
- (Jiang et al 2004) D. Jiang, C. Tang, and A. Zhang. "cluster analysis for gene expression data: A survey". IEEE Transactions on Knowledge

and Data Engineering, 16(11):1370–1386, 2004. doi: 10.1109/TKDE.2004.68.
- (Kodratoff and Michalski 1990) Y. Kodratoff and R. S. Michalski. "Machine Learning: An Artificial Intelligence Approach", volume 3. San Mateo, CA: Morgan Kaufmann, 1990.
- (Massa, et al. 2010) Massa J., Doorn J., Wainschenker R.; "Low-Coupled Parallel Strategy for Monte Carlo Radiation Dose Calculation"; Conf Proc IEEE Eng Med Biol Soc. 2010; pp. 1771 - 1774; ISSN: 1557-170X, Print ISBN: 978-1-4244-4123-5; DOI: 10.1109/IEMBS.2010.5626742; (2010).
- (Massa, et al. 2011) Massa J., Wainschenker R., Doorn J; "Optimización del cálculo por Monte Carlo de la Dosis en Radioterapia: Producción de Pares"; Congreso Argentino de Bioingeniería SABI 2011, Mar del Plata; (2011).
- (Perino, et al. 2012) Perino, Massa, D'Amato, Furlong; "Segmentación de imágenes médicas oritentado al cálculo de dosis por el método de Monte Carlo"; Asociación Física Argentina; Rio Cuarto, Córdoba; (2012)
- (Tristan et al 2009) Tristan P., Abrile P., Massa J., Ferraggine V., Rivero L. y Wainschenker R. Evolución en el desarrollo de SIG's hacia herramientas Open Source. ECImag 2009: 2da Escuela y Workshop de Ciencias de las Imágenes.
- (Xun, et al. 2010) Xun Jia; "Development of a GPU-based Monte Carlo dose calculation code for coupled electron–photon transport"; Phys. Med. Biol.; Vol. 55 pp. 3077; (2010).
- (Li, et al. 2013] Li, Y., Jiang, J., Zhang, M., & Wei, S. (2013, June). An Efficient Monte-Carlo Dose Calculation System for Radiotherapy Treatment Planning. InComputational and Information Sciences (ICCIS), 2013 Fifth International Conference on (pp. 314-317).
- (Fanti, et al. 2009) Fanti, V., Marzeddu, R., Pili, C., Randaccio, P., Siddhanta, S., Spiga, J., & Szostak, A. (2009, May). Dose calculation for radiotherapy treatment planning using Monte Carlo methods on FPGA based hardware. In Real Time Conference, 2009. RT'09. 16th IEEE-NPSS (pp. 415-419).
- (Luu, et al 2009) Luu, J., Redmond, K., Lo, W. C. Y., Chow, P., Lilge, L., & Rose, J. FPGA-based Monte Carlo computation of light absorption for photodynamic cancer therapy. In Field Programmable Custom Computing Machines, 2009. FCCM'09. 17th IEEE Symposium on (pp. 157-164).
- (Shi, et al. 2013) K. Shi, D. Boland, and G. Constantinides, "Accuracy-performance tradeoffs on an fpga through overclocking," Field-Programmable Custom. Computing Machines (FCCM), 2013 IEEE 21st Annual International Symposium on, 2013.
- (Boland, et al. 2013) D. Boland and G. Constantinides, "Revisiting the reduction circuit: a case study for simultaneous architecture and precision optimisation," Field-Programmable Technology (FPT), 2013 International Conference
- (Vazquez, et al 2014) Vazquez, M., & Tosini, M. (2014, November). Design and implementation of decimal fixed-point square root in LUT-6 FPGAs. In Programmable Logic (SPL), 2014 IX Southern Conference on (pp. 1-6). IEEE.
- (Buchmann, et al. 2014) Buchmann, J., Cabarcas, D., Göpfert, F., Hülsing, A., & Weiden, P. (2014). Discrete Ziggurat: A time-memory trade-off for sampling from a Gaussian distribution over the integers. In Selected Areas in Cryptography--SAC 2013 (pp. 402-417). Springer Berlin Heidelberg.
- (Arguiñarena, et al. 2014) "Visualización de imágenes médicas de alta resolución mediante una aplicación zero footprint", Arguiñarena E., del Fresno M., Massa J., Escobar P., Santiago M., Congreso Nacional de Ingeniería Informática/Sistemas de Información, CONAIISI 2014, 13-14 Noviembre de 2014, San Luis, Argentina, páginas 255-260, ISSN: 2346-9927.1.
- (Schmidhuber, et al. 2015) Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks, 61, 85-117.
- (Agrawal, 2014) Agrawal, P. (2014). Analysis of Multilayer Neural Networks for Object Recognition.
- (Bengio, et al. 2013) Bengio, Samy; Deng, Li; Larochelle, Hugo; Lee, Honglak; Salakhutdinov, Ruslan, "Guest Editors' Introduction: Special Section on Learning Deep Architectures," in Pattern Analysis and Machine Intelligence, IEEE Transactions on , vol.35, no.8, pp.1795-1797, Aug. 2013

FACILIDADES DISPONIBLES Y/O FORMA DE ACCESO Y FUENTES DE FINANCIAMIENTO

El equipamiento disponible incluye dos PC Pentium IV de 2,8 a 3,0 GHz, una PC Core i7 de 3.0 GHz y tres PC AMD de 3,0 GHz, cuatro de ellas cuentan con monitores de 22 pulgadas y todos con plaquetas de sonido, micrófonos y parlantes e integradas en red; además de una videocámara Sony Digital Handycam DCR-TRV520, dos impresoras láser blanco y negro dos impresoras láser color, dos bancos de discos Linksys NAS200 de 2 TB de capacidad cada uno, 2 Netbooks Asus EeePc 4G, un scanner AGFA.

Para la Línea 1 se dispone de una Base de casos de estudio propios y de UTN-FRBA, PUC-Rio, UNLP-Fac Informática, Grado y Postgrado UNLaM y el equipamiento específico para la Línea 2 incluye un cluster dedicado con 6 equipos de cálculo, dos Pentium IV Core 2 Duo de 2.2 GHz. con 2 Gb de RAM, uno Quad Core de 2.6 Ghz. Con 4 Gb de RAM, un AMD Athlon X4 con 2 Gb de Ram y 2 Core 2 Duo de 2.6 GHz. con 4 Gb. de RAM.

La Universidad provee acceso a la Biblioteca Electrónica y textos recientes en Biblioteca Central y se dispone de oficinas con una superficie de 70 m2, completamente equipadas.

Este proyecto será financiado con el subsidio que la Secretaría de Ciencia y Técnica de la UNICEN asigna al Instituto INTIA. Las comunicaciones a congresos y publicaciones serán financiadas parcialmente con recursos provenientes de las partidas del Departamento de Computación y Sistemas de la Facultad de Ciencias Exactas de la UNICEN.