Memoria Organizacional Retroalimentada por Flujos de Datos Aplicada a la Predicción de Pasturas

Mario José Diván
Facultad de Ingeniería
Universidad Nacional de La Pampa
General Pico, CP 6360, Argentina
mjdivan@ing.unlpam.edu.ar

María de los Ángeles Martín
Facultad de Ingeniería
Universidad Nacional de La Pampa
General Pico, CP 6360, Argentina
martinma@ing.unlpam.edu.ar

Abstract

Se presenta aquí, un resumen del diseño de memoria organizacional basada en casos, que permite el intercambio de conocimiento en el contexto de una arquitectura de procesamiento de flujos de datos basada en techología Big Data. La memoria organizacional recomienda cursos de acción al proceso de toma de decisiones en línea de la arquitectura, y a su vez, el tomador de decisiones de la arquitectura retroalimenta la memoria organizacional a partir de las decisiones tomadas y acciones realizadas.

Un aspecto clave asociado con el procesamiento de flujos de datos es que las mediciones deben ser consistentes y comparables en cualquier momento para tomar decisiones correctamente. De esta manera, la arquitectura de procesamiento se basa en el marco C-INCAMI (Context-Information Need, Concept Model, Attribute, Metrics and Indicator) para definir los proyectos de medición y evaluación (M&E). Así, este trabajo expone la relación entre el marco de M&E, la arquitectura de procesamiento de flujos de datos sustentada en Big Data y la memoria organizacional. Con el fin de ilustrar su utilidad, un caso práctico que emplea datos del radar meteorológico (RM) de la Estación Experimental Agrícola (EEA) de INTA Anguil es planteado. En este caso, se presenta un ejemplo de sistema de recomendación que consiste en la predicción de pasturas para la producción agricola.

Palabras Clave— Memoria Organizacional, Recomendación, Medición, Grandes Datos, Flujo de Datos.

1. Introducción

Actualmente existen aplicaciones que procesan un conjunto de datos en forma continua, y ante cada arribo [1, 2]. Estas arquitecturas pueden definir dinámicamente la topología de procesamiento sobre los flujos de datos, ajustándose a diferentes necesidades de cómputo, y delegando la definición estructural y significado del dato en la lógica embebida dentro de la aplicación. En este tipo de soluciones se categoriza el Enfoque Integrado de Procesamiento de Flujos de Datos centrado en Metadatos de Mediciones (EIPFDcMM) [3], el cual sustentado en el marco de medición y evaluación C-INCAMI (Context-Information Need, Concept model, Attribute, Metric and Indicator), incorpora metadatos al proceso de medición, promoviendo la repetitividad, comparabilidad y consistencia del mismo. Desde el punto de vista del sustento semántico y formal para la medición y evaluación (M&E), C-INCAMI establece una ontología que incluye los conceptos y relaciones necesarias para especificar los datos y metadatos de cualquier proyecto de M&E. Por otra parte, y a diferencia de otras estrategias de procesamiento de flujos de datos [4, 5, 6, 7], gracias a la incorporación de metadatos, el EIPFDcMM es capaz de guiar el procesamiento de las medidas, analizando cada una en base al significado definido en el proyecto de M&E y dentro de su contexto de procedencia. Adicionalmente, ello permite incorporar un comportamiento detectivo y predictivo sobre las medidas contextualizadas, lo que posibilita un monitoreo activo sobre las entidades bajo análisis sustentado en una memoria organizacional [8] capaz de gestionar experiencias previas y soportar el proceso de toma de decisiones.

La Arquitectura de Procesamiento centrada en Metadatos de Mediciones (APcMM), evoluciona la estrategia original [9] con la incorporación de un repositorio Big Data en el contexto de computación distribuida. Esto implica la necesidad de reunir las tecnologías Big Data y el procesamiento de flujos de datos; lo que garantiza una mayor capacidad de procesamiento ante grandes volúmenes de datos, lo que posibilita una gestión eficiente del conocimiento en la memoria organizacional.

La memoria organizacional que integra la Arquitectura de Procesamiento centrada en Metadatos de Mediciones sirve como base para el intercambio de conocimiento de la organización, y adicionalmente, para ser utilizada en sistemas de recomendación en los procesos de toma de decisiones. La gestión del conocimiento organizacional representa un activo clave como apoyo a los procesos de toma de decisiones por parte de diferentes grupos de interés de la organización. El objetivo principal de los sistemas de gestión del conocimiento consiste en gestionar, almacenar y recuperar el conocimiento de la organización, de modo que pueda ser utilizado más tarde para aprender, compartir conocimientos, resolver problemas, y en última instancia, para una mejor toma de decisiones.

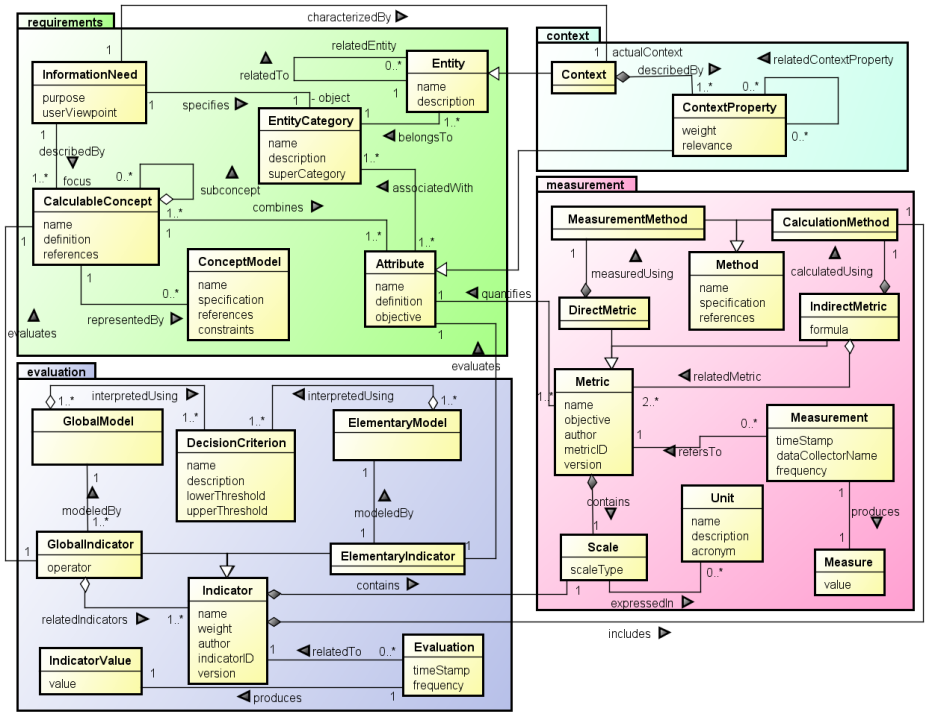

Figura 1. Principales componentes, conceptos y relaciones de C-INCAMI

Por lo tanto, con el fin de alcanzar y mantener la eficacia y competitividad de empresa, la organización necesita aprender de las experiencias del pasado y presente. En tal sentido, es necesario formalizar memorias organizacionales para hacer explícito el conocimiento tácito de los individuos -y por qué no- el conocimiento tácito de la comunidad también. De este modo, uno de los principales objetivos de la gestión del conocimiento de la organización, es hacer explícito el conocimiento implícito de los individuos, con el objetivo de formalizar el conocimiento informal a los efectos de permitir inferencias semánticas procesables por una máquina. En esta línea, una manera de resolver este problema, desde el punto de vista de la representación del conocimiento, es almacenando el conocimiento de una manera estructurada y formal. Nosotros hemos seguido este enfoque mediante el uso de la estrategia de memoria organizacional basada en casos, que combina las tecnologías de gestión del conocimiento con el razonamiento basado en casos (CBR) para representar cada elemento de conocimiento informal. La estructuración de una memoria organizacional en casos, facilita la captura, recuperación, transferencia y reutilización de conocimiento para la resolución de problemas en forma automática.

La principal contribucion de esta investigación es la integración de la memoria organizacional y el razonamiento basado en casos en la APcMM y la descripción de cómo dicha memoria organizacional puede ser usada para la toma de deciciones. Además, se discute cómo los datos, la información y el conocimiento, obtenidos a partir de fuentes de datos heterogéneas y distribuidas, pueden ser automáticamente y semánticamente procesables por las aplicaciones, por ejemplo, un sistema de recomendación "inteligente" para apoyar un proceso de toma de decisiones más eficaz. Con el fin de ilustrar el uso de la memoria organizacional, en un caso práctico relacionado con el radar meteorológico (WR) de la Estación Experimental Agrícola (EEA) INTA Anguil, se muestra un ejemplo de sistema de recomendación que consiste en la predicción de pasturas para la producción agrícola. También se describen las tendencias futuras y observaciones finales.

Este artículo está organizado en siete secciones. La Sección 2, resume el marco conceptual C-INCAMI y presenta C-INCAMI/MIS (*Measurement Interchange Schema*) para el intercambio de medidas que permiten retroalimentar la memoria organizacional. La sección 3 describe la arquitectura de procesamiento basado Metadatos de Mediciones. La sección 4 describe la memoria organizacional basada en casos. La Sección 5 ilustra la aplicación de la memoria organizacional a un caso práctico: un sistema de predictor de pasturas utilizando los datos del radar meteorológico de EEA INTA Anguil. La Sección 6 discute trabajos relacionados, y finalmente se resumen las conclusiones y trabajos futuros.

2. Panorama de C-INCAMI

C-INCAMI es un marco conceptual [10, 11] que define los módulos, conceptos y relaciones que intervienen en el área de M&E, para organizaciones de software (Ver figura 1). Se basa en un enfoque en el cual la especificación de requerimientos, la medición y evaluación de entidades y la posterior interpretación de los resultados están orientadas a satisfacer una necesidad de información particular. Está integrado por los siguientes componentes principales: 1) *Gestión de Proyectos de M&E*; 2) *Especificación de Requerimientos no Funcionales*; 3) *Especificación del Contexto del Proyecto*; 4)

Diseño y Ejecución de la Medición; y 5) *Diseño y Ejecución de la Evaluación*. La mayoría de los componentes están soportados por los términos ontológicos definidos en [11].

El componente *Gestión de Proyectos de M&E* (no expuesto en la Figura 1), define y relaciona un conjunto de términos de proyecto necesarios para lidiar con las actividades de M&E, los métodos, roles y artefactos.

El componente de *Especificación de Requerimientos No Funcionales* (paquete *requirements* en la Figura 1) permite especificar la necesidad de información (*Information Need* en Figura 1) de cualquier proyecto de M&E. La necesidad de información identifica el propósito (por ejemplo, entender, predecir, monitorear, etc) y el punto de vista del usuario (por ejemplo, usuario final). Adicionalmente, se focaliza sobre un concepto calculable (por ejemplo, calidad del sistema, calidad de los datos, etc) y es posible especificar la categoría de la entidad a evaluar (por ejemplo, el sistema, un recurso, etc). El concepto calculable es una relación abstracta entre los atributos de una entidad y una necesidad de información dada, la cual puede ser representada mediante un modelo conceptual. En este modelo conceptual, las hojas del modelo empleado son atributos, los cuales pueden ser cuantificados a partir de las métricas.

La componente *Especificación del contexto del proyecto* (*Context* en Figura 1) permite representar el estado de situación de la entidad que está siendo monitoreada. Se considera al contexto como un tipo especial de entidad en la que entidades relevantes están involucradas. Así, para describir el contexto se emplean los atributos asociados con las entidades relevantes, denominados propiedades de contexto, los cuales pueden ser cuantificados a través de las métricas al igual que los atributos de la entidad bajo análisis. De este modo, es posible contar con una perspectiva cuantitativa no solo de la entidad bajo análisis, sino también de su contexto caracterizado a partir de sus propiedades contextuales.

El componente *Diseño y Ejecución de la Medición* (*Measurement* en Figura 1), incluye los conceptos y relaciones necesarias para especificar el diseño e implementación de la medición. De este modo, el diseño de la medición la métrica brinda una especificación de la medición indicando cómo cuantificar un atributo particular de la entidad bajo análisis, empleando para ello un método, e indicando cómo representar su valor utilizando una escala en particular. Pueden distinguirse dos tipos de métricas: Directas e Indirectas. Por un lado, se denominan métricas directas a aquellas en las que los valores son obtenidos directamente desde la medición del atributo correspondiente a la entidad. Por otro lado, se denominan métricas indirectas a aquellas cuyo valor es calculado o derivado, a partir de los valores de otras métricas siguiendo alguna fórmula. Así, la ejecución de la medición refiere a las tareas involucradas para efectivizar la obtención de la medida (el valor) a partir de la métrica definida. En este sentido, debemos mencionar que C-INCAMI supone que los valores obtenidos son siempre deterministas, es decir, que ante una medición se obtiene siempre una medida. Este supuesto se torna endeble ante situaciones como la planteada por Abajo Martínez en [12] en donde el tiempo o recurso involucrado para obtener un valor determinista, puede incluso ser hasta perjudicial en una actividad económica, prefiriéndose emplear distribuciones de probabilidad para mejorar el rendimiento de la línea de producción monitoreada en tiempo real. A partir de ello, es que hemos incorporado este aspecto, entre otras cuestiones, como extensiones conceptuales de C-INCAMI [13], para poder lidiar con distribuciones de probabilidad y grupos de seguimiento a partir de las fuentes de datos asociadas con las métricas directas.

El *Diseño y Ejecución de la Evaluación* (*Evaluation* en figura 1), incluye los conceptos y relaciones necesarias para especificar el diseño e implementación de la evaluación. La evaluación se nutre de las métricas definidas, e interpreta las medidas a través del concepto del indicador. Existen dos tipos de indicados: Elementales y Globales. Por un lado, el indicador elemental evalúa los atributos combinados en un modelo conceptual, obteniendo su valor a partir de una función de mapeo desde las medidas de las métricas que cuantifican los atributos. De este modo, el valor del indicador es interpretado a través de los criterios de decisión, los cuales permiten analizar el nivel de satisfacción logrado en términos de los requerimientos no funcionales. Tales criterios de decisión, son provistos por expertos del negocio bajo análisis y retroalimentados a partir de la experiencia y conocmiento formalizado de la memoria organizacional. Por otro lado, los indicadores globales representan el nivel de satisfacción general en términos de la necesidad de información planteada, nutriéndose y relacionando los diferentes indicadores elementales.

Los flujos de medidas que se informan desde las fuentes de datos al APcMM, se estructuran incorporando a las medidas, metadatos basados en C-INCAMI con las extensiones planteadas en [13], para poder gestionar no solo información sobre la métrica a la que corresponde y el atributo de la entidad que se mide, sino también poder incorporar el grupo de seguimiento asociado, si la medida es determinista o no con su respectiva probabilidad, entre otros aspectos.

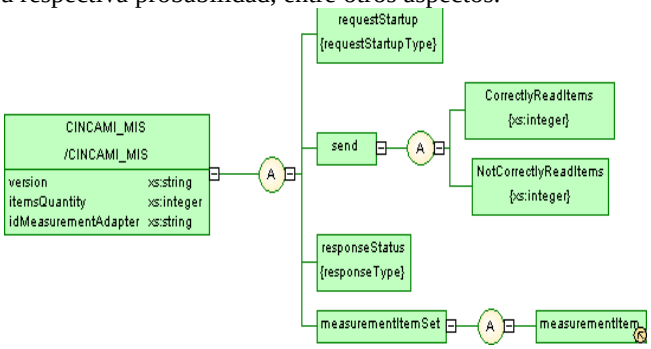

Figura 2. Nivel superior del esquema C-INCAMI/MIS

En tal sentido, C-INCAMI/MIS (Ver Figura 2) es el esquema de intercambio de mediciones que permite dentro de un mismo flujo de datos etiquetar conjuntamente con cada medida asociada al atributo, las medidas vinculadas a cada propiedad de contexto. El conjunto de mediciones del flujo se organiza bajo la etiqueta denominada *measurementItemSet* de la Figura 2, e identificando bajo cada etiqueta

measurementItem, a una medida con sus respectivas propiedades de contexto. Esto representa un aspecto importante en APcMM para la gestión de mediciones ya que, al disponer de fuentes heterogéneas de datos, es posible homogenizar las mediciones bajo un mismo esquema independientemente del origen mediante el Adaptador de Mediciones (Ver sección 3). Así, tanto el almacenamiento, como el procesamiento y los servicios a terceros, gestionarán siempre flujos C-INCAMI/MIS sin importar la fuente que los haya generado, lo que facilita su consulta, intercambio y extensibilidad, principalmente teniendo en cuenta la retroalimentación de la memoria organizacional.

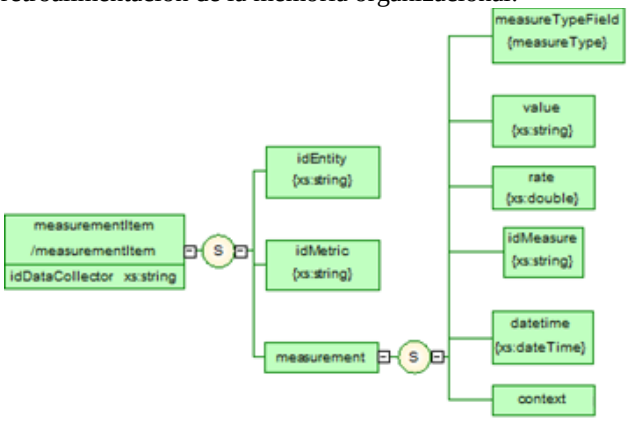

Figura 3. Estructura de measurementItem en C-INCAMI/MIS

Como puede apreciar en la figura 3, cada etiqueta *measurementItem* tiene asociada la identificación del recolector de datos (*idDataCollector*) y representa una medición. Esta es informada conjuntamente con el identificador de la entidad bajo análisis (*idEntity*) y la métrica (*idMetric*) responsable de la cuantificación del atributo en términos del proyecto de M&E. La información particular a la medición se encuentra bajo la etiqueta *measurement*, además de contar con la fecha y hora (*datetime*) en la cual se obtuvo la medida, cuenta con el atributo *measureTypeFiled* el cual puede ser ACTUAL o ESTIMATED. Si el mencionado atributo es ACTUAL implica que el valor informado es determinista, pero si fuere ESTIMATED, la medida es no determinista motivo por el cual se informará cada par (valor -*value*-; probabilidad -*rate*-) bajo un mismo valor *idMeasure*, lo cual permite agrupar a la distribución de probabilidad.

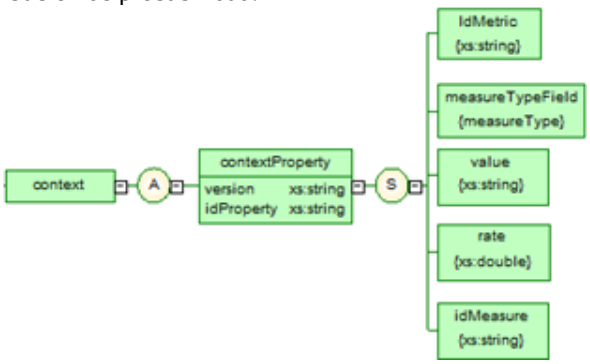

Figura 4. Estructura de la etiqueta Context en C-INCAMI/MIS

Adicionalmente, es posible cuantificar el contexto de la entidad bajo análisis a partir de sus propiedades contextuales. Esto es lo que representa la etiqueta *context* en la figura 3 y detalla la figura 4.

Como puede apreciarse en la figura 4, bajo la etiqueta *Context* se incorporan todas las propiedades de contexto (*ContextProperty*) que son cuantificadas al mismo instante en que lo es la medida del atributo para la entidad bajo análisis (Ver figura 3). Al igual que la métrica asociada con un atributo de la entidad bajo análisis, cada métrica asociada con una propiedad contextual puede arrojar un valor determinista o no detemrinista, motivo por el cual el concepto asociado con las etiquetas *measureTypeField*, *value*, *rate* y *idMeasure* es análogo al anteriormente expresado solo que circunscripto a la propiedad contextual.

De este modo, tanto C-INCAMI como C-INCAMI/MIS representan aspectos esenciales para APcMM por cuanto fomenta la interoperabilidad, a la vez que el procesamiento de los flujos de datos puede efectivamente ser guiado a partir de los metadatos embebidos en el flujo. Adicionalmente, la posibilidad de homogenizar el intercambio y almacenamiento de las medidas asociadas con un proyecto de M&E, permite retroalimentar la Memoria Organizacional desde el ajuste de los criterios de decisión asociados con los indicadores, hasta la captación de nueva experiencia que permita expandir el conocimiento organizacional prexistente.

3. Arquitectura de Procesamiento centrada en Metadatos de Medición

La APcMM [14, 15, 16] es una estrategia de procesamiento de flujos de datos especializada en proyectos M&E y sustentada en C-INCAMI [10, 11], la cual incorpora comportamiento detectivo y predictivo en línea, a la vez que permite el aprovisionamiento a terceros de los flujos mediante suscripción, y el almacenamiento de las medidas en grandes repositorios para responder consultas de datos ad-hoc guiado por una Memoria Organizacional basada en casos, a introducir en la sección 4.

Sintéticamente y como puede apreciarse en la figura 5, la idea conceptual de procesamiento consiste en que los flujos de medidas provienen desde fluentes de datos heterogéneas (por ejemplo, un radar) estructurados bajo el esquema C-INCAMI/MIS. Cada flujo C-INCAMI/MIS es generado a partir de la fuente de datos por un adaptador de mediciones (*MA* en figura 2) que establece la correspondencia entre la medida, sus metadatos y las propiedades de contexto en base al proyecto de M&E definido. Así, cada flujo C-INCAMI/MIS es enviado desde el MA a la función de reunión informando las medidas, sus propiedades de contexto y sus metadatos. De este modo, la función de reunión: a) incorpora el flujo en el repositorio de grandes datos, b) provee el flujo en tiempo real a los terceros suscriptos al servicio, y c) provee una copia del flujo a la función de análisis y suavización. Esta última, realiza diversos análisis estadísticos (por ejemplo, análisis de correlación) sobre las métricas del flujo, permitiendo almacenar una instantánea de la situación de la entidad bajo análisis en memoria, disparar alarmas en caso de desvíos respecto de los establecido en el proyecto de M&E al tomador de decisiones (*Decision Maker* –DM- en Figura 2), y suavizar

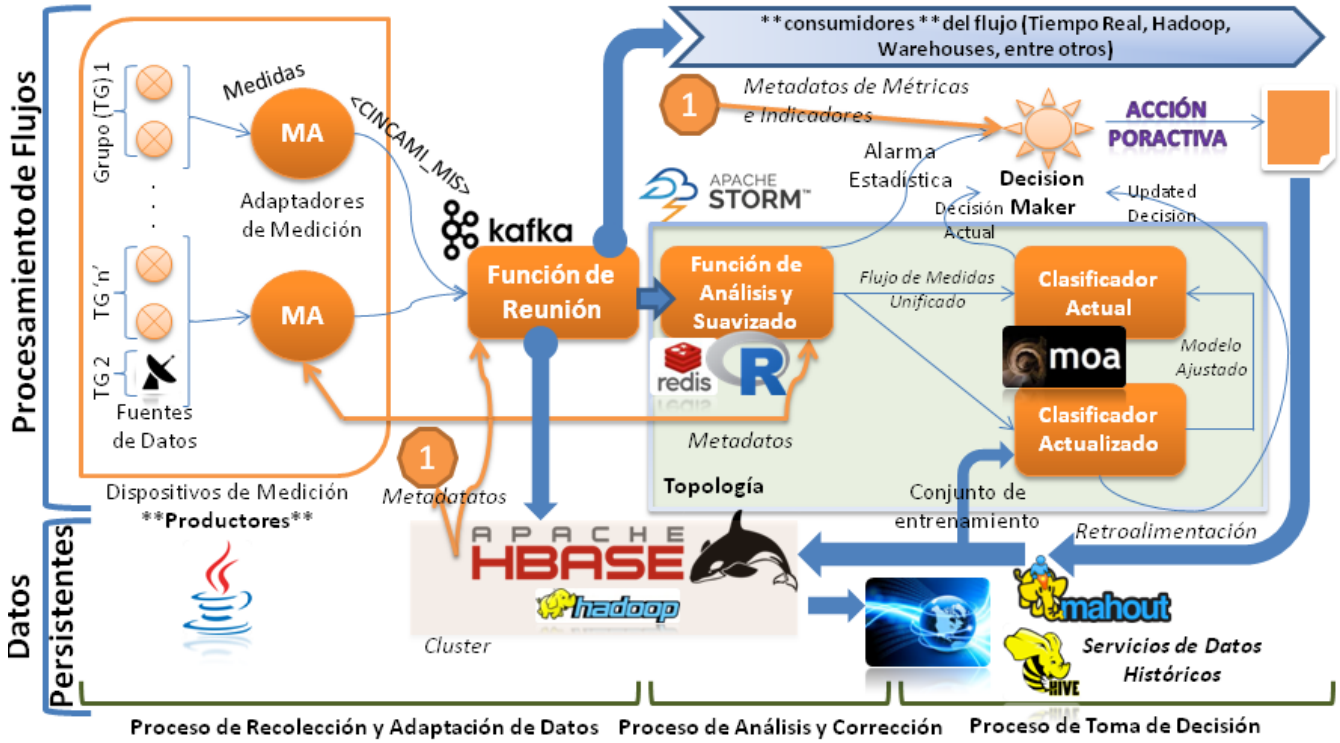

Figura 5. Arquitectura de Procesamiento centrada en Metadatos de Mediciones con Big Data

el flujo en base a la configuración del proyecto de M&E (por ejemplo, filtrar valores atípicos). De este modo, los flujos suavizados se informan al clasificador actual, quien: a) Toma una decisión al instante (D^t), y b) En paralelo, se actualiza incrementalmente, generando un nuevo clasificador actualizado, y toma una nueva decisión (D^t+1). Si algunas de las decisiones Dt o Dt+1 se corresponden con una situación de eventual riesgo según lo definido en el proyecto de M&E, se dispara una alarma al DM. Ambos modelos, el clasificador actual y el actualizado, son comparados en línea contrastando su área bajo la curva ROC (acrónimo de *Receiver Operating Characteristic*) [17], y aquel que mayor área bajo la curva posea se tornará en el nuevo clasificador actual. De este modo, el clasificador no sólo aprende desde el conjunto de entrenamiento dado por la memoria organizacional del repositorio en el momento cero, sino que va ajustando incrementalmente su comportamiento a partir de los datos tratados estadísticamente en línea para ajustarse a nuevas situaciones, retroalimentado a su vez a la memoria organizacional. Un mayor detalle puede encontrarse en [14, 15], donde los procesos han sido formalizados mediante el metamodelo SPEM [18] para promover su comunicabilidad y extensibilidad.

Desde el punto de vista técnico, la APcMM evoluciona la solución propuesta en el EIPFDcMM, ya que ahora es posible la gestión de repositorios de grandes datos en entornos de computación distribuida, junto la provisión de datos mediante servicios por suscripción, en forma adicional al procesamiento de flujos original. Así y a los efectos de promover la extensibilidad, dinamismo y difusión de la arquitectura, se ha priorizado el empleo de tecnologías de código abierto, maduras y escalables.

De este modo, las fuentes continuan implementando la interface DataSource original, a través de la cual se definen las responsabilidades que una fuente del EIPFDcMM debe satisfacer para aprovisionar datos a la arquitectura, pero ahora en APcMM se constituyen adicionalmente en productoras en términos de Apache Kafka [19, 20] como puede apreciarse en la Figura 5. Así, los datos enviados por los productores (por ejemplo, un radar), serán procesados por un servicio de suscripción dentro del cluster de procesamiento de mensajería, bajo el concepto de función de reunión, donde lógicamente todas las medidas de la misma entidad bajo análisis, son agrupadas para informarse en forma conjunta mediante esta misma tecnología a los consumidores. Además, los consumidores podrán procesar en tiempo real el flujo C-INCAMI/MIS reunido a partir de Kafka, entendiéndose por tales a: 1) Los suscriptores que consumen en tiempo real el flujo de medidas a partir de Apache Kafka, 2) La topología de procesamiento de flujos de datos sustentada en Apache Storm [21] que continuará con el procesamiento en tiempo real, y 3) Apache HBase [22, 23] como repositorio de grandes datos que almacena las medidas para su uso posterior.

Así, la estrategia interna de procesamiento de flujos de datos se monta ahora sobre Apache Storm [21], y consume los flujos en forma continua desde la función de reunión a partir de Apache Kafka, como puede apreciarse en el recuadro de la Figura 5 denominado "Topología". Esto último, aporta flexibilidad, escalabilidad y dinamismo respecto de la

configuración de las topologías[1] de procesamiento de datos, ya que tanto la función de suavización como los clasificadores se corresponden con Bolts1 que pueden ser reorganizados en forma ágil y simple dentro de la misma. Adicionalmente, dentro de la topología de procesamiento ejecutada sobre Apache Storm, APcMM continua empleando R [24] para los cómputos estadísticos de la función de análisis y suavizado, empleando a partir de ahora Redis [25] como base de datos NoSQL en memoria para: i) Gestión de cache, ii) La utilización de resultados intermedios desde R, y iii) El almacenamiento de las instantáneas sobre el ultimo estado conocido de cada entidad bajo análisis. Finalmente, se utiliza dentro de la topología de procesamiento los clasificadores del marco Massive Online Analysis (MOA) [26] que permiten actualizaciones incrementales a la vez que están nativamente preparados para minería de flujos.

Por otro lado, y en relación a la gestión persistente de medidas, la arquitectura emplea para el almacenamiento y procesamiento de grandes datos un cluster Apache Hadoop [27] para promover el procesamiento distribuido, con una base de datos columnar Apache HBase [22, 23] que permite el escalamiento monolítico y el acceso aleatorio a las mediciones asociadas con un proyecto de M&E dado. A partir de este repositorio de medidas provenientes desde diferentes orígenes, se emplea Apache Hive [28] para soportar consultas ad-hoc sobre entornos de cómputo distribuido, al igual que Apache Mahout [29] para poder llevar adelante diferentes análisis de agrupamiento y clasificación que permitan detectar nuevos patrones de comportamientos respecto del objetivo del Proyecto de M&E. Esto posibilita que la Memoria Organizacional tome ventajas respecto de las capacidades de paralelismo, distribución y escalabilidad que ofrece este nuevo contexto de procesamiento, ya que el motor de razonamiento basado en casos de la memoria organizacional, utiliza programas tipo MapReduce [30] en su sistema de recomendación, para estructurar naturalmente cada conjunto (hechos, solución) como <clave, valor>.

De este modo, la arquitectura técnicamente no solo se orienta al cómputo distribuido, su escalabilidad y extensibilidad en base a productos maduros con el objetivo de poder enfrentar repositorios de grandes datos, sino que también ahora se nutre de Apache Storm, para dotar a la topología de procesamiento de dinamismo, facilitando su versionado e interoperabilidad ante eventuales cambios de requerimientos.

4. Memoria Organizacional Basada en Casos

Una vez que los flujos C-INCAMI/MIS son incorporados desde las fuentes de datos al repositorio persistente de grandes datos en APcMM, es conveniente estructurar los mismos en una memoria organizacional, de manera que posteriormente pueda ser explotada y utilizada para la recomendación durante el proceso de toma de decisión.

El conocimiento aporta ventaja estratégica en materia de competitividad empresarial, en tal sentido, los sistemas de administración del conocimiento permiten administrar y almacenar el conocimiento organizacional, con el objetivo de ser utilizado para aprender, resolver problemas y como apoyo a la toma de decisiones [31]. Nuestra propuesta, es almacenar el conocimiento aportado por los flujos de datos y sus metadatos, en forma estructurada bajo una Memoria Organizacional Basada en Casos [32].

Un caso es una pieza contextualizada de conocimiento que representa una experiencia. Contiene la lección pasada que es el contenido del caso y el contexto en el cual la lección puede ser utilizada [33]. Típicamente, un caso comprende:

- El problema que describe el estado del mundo cuando ocurrió el caso
- La solución que describe cómo se resuelve el problema, y/o
- El resultado que describe el resultado obtenido como consecuencia de la solución del problema.

El proceso de razonamiento basado en casos consiste en asignar valores a las variables de características del problema (caracterizar el problema), y encontrar los valores adecuados para las instancias de la solución, a través de criterios de evaluación de similitud de casos. Tales características tienen su analogía con los atributos que permiten describir la entidad bajo análisis mediante C-INCAMI, y cuyas medidas son informadas a través del flujo C-INCAMI/MIS.

Tradicionalmente, hay varios tipos de métodos para representar casos, que van desde representaciones no estructuradas a totalmente formales y automáticamente procesables [34]. Estos últimos, están basados en representaciones totalmente estructuradas, y consisten en aplicar, por ejemplo, técnicas orientadas a objeto centradas en el uso de metadatos.

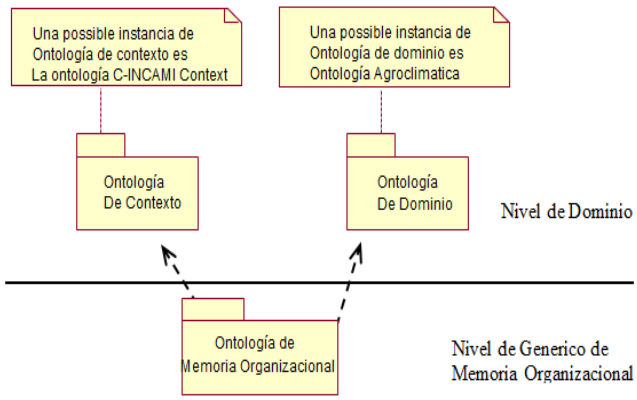

Figura 6. Relación entre ontologías de dominio y la ontología específica de Memoria Organizacional

[1] Se entiende por *Topología* en Apache Storm a un conjunto de fuentes de datos (denominadas Spout) que proveen datos a uno o más componentes vinculados (denominados Bolt), a partir de los cuales se realiza alguna síntesis, transformación o disgregación del flujo de datos original a los efectos de ser consumido por un usuario final, o bien, constituir la entrada de uno o más componentes (otros Bolts) [4].

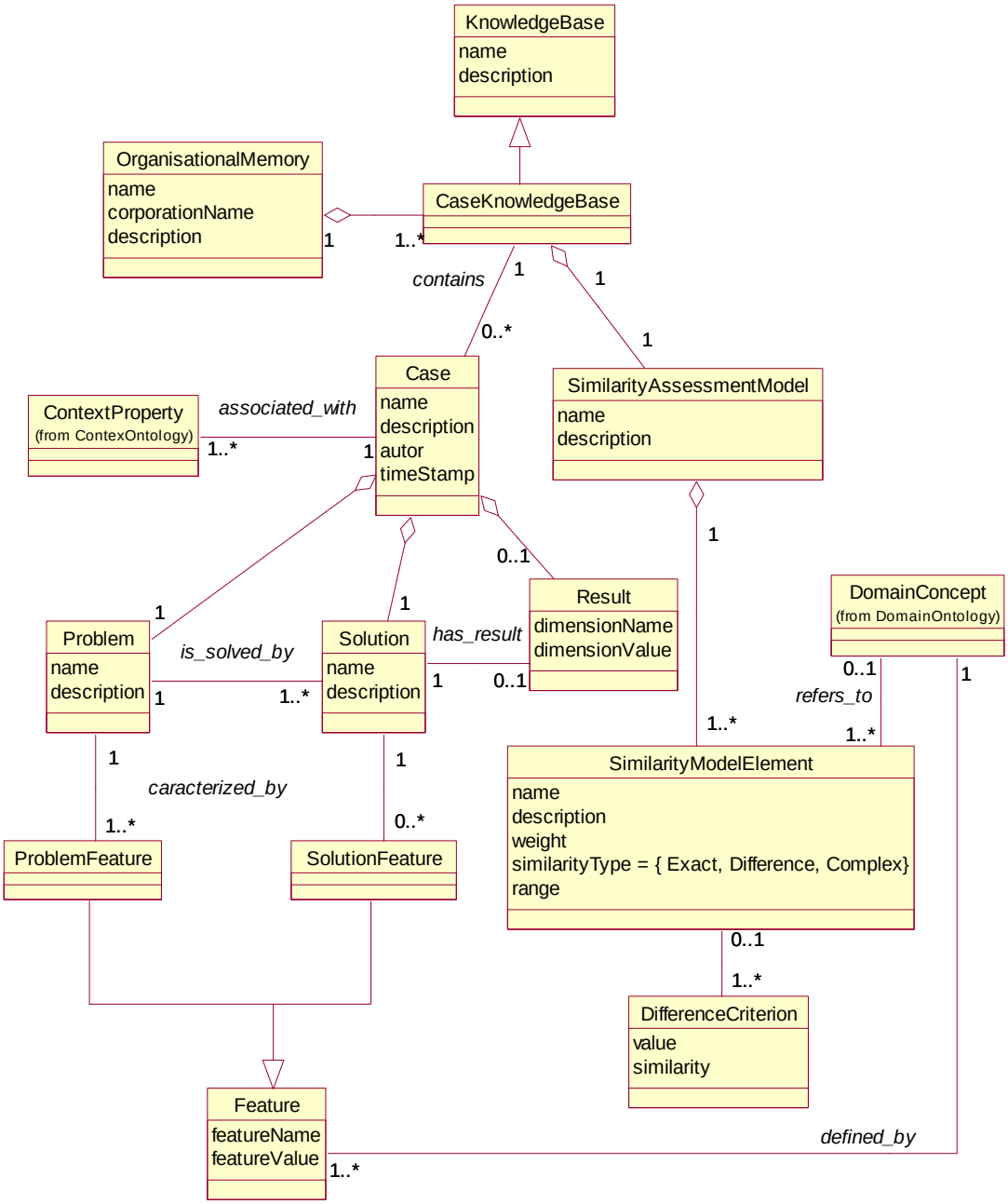

Figura 7. Modelo conceptual de la ontología de memoria organizacional basada en casos

La memoria organizacional propuesta está basada en ontologías [31] que operan en dos niveles distintos de abstracción (Ver figura 6): Por un lado, en el nivel de memoria organizacional genérico, se define la ontología de memoria organizacional en sí; y por otro lado, para caracterizar los casos de acuerdo al dominio de conocimiento y teniendo en cuenta su contexto [10], se necesita proveer al marco con ontologías de dominio y contexto respectivamente (ontologías de nivel de dominio).

A seguir, la ontología de memoria organizacional es resumida en la sección 4.1, la representación del modelo de similitud es introducida en la sección 4.2, y una medida de similitud es sintetizada en 4.3.

4.1. Ontología de Memoria Organizacional Basada en Casos

Si bien los beneficios del uso de los sistemas de gestión del conocimiento son bien conocidos, y la idea de aplicar los métodos de razonamiento basado en casos a lecciones aprendidas y buenas prácticas no son nuevas en el área de representación del conocimiento, casi no hay consenso todavía sobre muchos de los conceptos y terminología utilizada tanto en la gestión del conocimiento como en el área de razonamiento basado en casos. Con el fin de alcanzar este objetivo hemos construido una conceptualización común para

la memoria organizacional basada en casos donde los conceptos, atributos y sus relaciones se especifican de forma explícita; lo que constituye uno de los pasos básicos para la construcción de la ontología.

En esta sección se describen los principales conceptos de la ontología de memoria organizacional basada en casos [32], que se ilustran en el diagrama UML de la fugura 7.

Una memoria organizacional basada en casos, es un repositorio que almacena el conocimiento adquirido en experiencias pasadas como casos, esto es, como lecciones aprendidas, buenas prácticas, heurísticas, etc. Para una mejor organización y búsqueda de dichas experiencias, la memoria organizacional se compone de varias bases de conocimientos basada en casos (*CaseKnowledgeBase* en figura 7), que agrupan los casos por conocimientos de distintas áreas.

Un caso es una pieza contextualizada de conocimiento que representa una experiencia, por lo que es fundamental en toda memoria organizacional guarda la información del contexto donde ocurre cada caso. Por lo tanto, a cada caso se le asocian las propiedades de contexto (*Context Properties* en figura 7) correspondientes al dominio de aplicación, definidas en la ontología de contexto.

La representación del conocimiento a través de casos, facilita el reúso del conocimiento adquirido en situaciones de problemas similares pasados para ser aplicado a un nuevo problema [35]. En una definición formal un caso es un par ordenado <*P*, *S*>, donde *P* representa el espacio del problema, mientras que *S* se asocia con el espacio de la solución.

Los problemas y las soluciones se describen a través de variables de características del problema (*ProblemFeature* en figura 7) y variables de características de la solución (*SolutionFeature* en figura 7) respectivamente. El proceso de razonamiento basado en casos, consiste en asignar valores a las variables características del problema, y encontrar los valores adecuados para las instancias de la solución, a través de criterios de evaluación de similitud de casos (En la sección 4.2 se muestra un modo de cálculo para la similitud de casos). Por lo tanto, en cada tipo de conocimiento debe especificarse un modelo de similitud (*SimilarityAssessmentModel* en figura 7).

Para que una memoria organizacional pueda ser implementada en la web semántica, y pueda ser procesada automáticamente, necesita tener asociada una ontología de dominio [31], la cual proporciona la terminología (*Domain Concept* en figura 7) que provee los tipos de las variables que caracterizan al problema y a la solución.

4.2. Representación del Modelo de Similitud

Para que un sistema CBR (*Case-Based Reasoning*) sea útil a una organización, debería ajustarse a las principales fuentes de conocimiento de la empresa, y por lo tanto necesitan funciones de similitud apropiadas a cada base de casos [33]. En esta sección se propone un modelo que permite definir la estructura de un caso indicando las características (*feature* en figura 7*)* que lo describen y su modelo de similitud. Por ejemplo, para el dominio de la producción de pasturas, una base de casos podría guardar conocimiento relacionado con la caracterización de las regiones productivas, y otra al tipo de pastura y rindes estimados asociados, que sirva como base para recomendar tipos pasturas en función de las actividades productivas en particular. La forma en que se caracterizan y se evalúan la similitud de los casos de las regiones productivas, es completamente distinta a como se lo hace para un tipo de pastura en particular, siendo necesario, por lo tanto, definir la estructura del caso y el modelo de similitud apropiado a cada base de casos.

Como se observa en el modelo de la figura 7, a cada base de conocimiento basado en casos se le asocia un modelo de similitud (*SimilarityAssessmentModel*), que se compone de varios elementos de similitud (*SimilarityModelElement*), uno para cada característica constituyente del caso. De este modo, el modelo conceptual expuesto en la figura 7, define la estructura de la memoria organizacional a partir de la cual se entrenarán los clasificadores del APcMM (Memoria Organizacional implementada mediante Apache HBase y Hadoop en Figura 5), como así también se procederá a retroalimentar la estrategia mediante las decisiones generadas por el tomador de decisiones del mismo.

4.3. Medida de Similitud

Tradicionalmente, la similitud entre un caso recuperado *R* y un nuevo caso *C*, se define como la suma de las similitudes entre los valores de sus características constituyentes multiplicados por sus pesos de relevancia relativa:

$$Similitud(R, C) = \sum_{f \in F} w_f . sim_f(f_R, f_C)$$

En donde w_f es el peso de relevancia de la característica *f* y sim_f es la función de medida de similitud de una característica especifica *f*, perteneciente al conjunto *F* de todas las características disponibles.

Por lo tanto, para proveer una representación adecuada de la similitud, es necesario representar tanto los pesos de relevancia como la descripción de la función de similitud para una característica específica. Los pesos se representan como un atributo dentro de cada elemento de similitud, y la función de similitud se restringe a tres tipos generales de funciones de similitud: *Exact*, *Difference y Complex* [35, 36].

- La función de similitud *Exact*, devuelve 1 si los valores de característica son iguales, y 0 en otro caso.
- La función de similitud *Difference*, es inversamente proporcional a la diferencia entre los valores de las características. Esta función solamente se puede aplicar cuando es posible definir la diferencia entre los valores.
- La función de similitud *Complex*, resuelve la similitud para todas aquellas situaciones donde las dos funciones de similitud anteriores no son aplicables.

En nuestro modelo, estos parámetros están representados en la clase *DifferenceCriterion.*

5. Un Caso Práctico

Con el fin de ilustrar los conceptos, atributos y relaciones definidos anteriormente, vamos a elaborar en un ejemplo de base de conocimiento basado en casos y su modelo de evaluación de similitud para un dominio específico: un sistema de predicción de pasturas, utilizando los datos del radar meteorológico de EEA INTA Anguil. Esta base de casos almacena un conjunto de conocimientos relacionados con el crecimiento del pasto en base a una serie de datos, incluyendo las condiciones actuales y pronósticos meteorológicos, eventos de lluvia y los registros climáticos del pasado, procesados por la arquitectura APcMM y teniendo el radar como fuente de datos. De este modo, la sección V.A presenta una síntesis de los datos provistos por el radar de INTA, mientras que la sección V.B introduce la base de conocimiento.

5.1. RADAR de la EEA INTA Anguil

La Estación Experimental Agropecuaria (EEA) INTA Anguil tiene instalado un Radar Meteorologico (RM) marca Gematronik modelo Meteor 600C (Ver figura 8.a) que genera un flujo de datos estimado de 17Gb diarios, lo que representa un desafío para su almacenamiento, gestión y posterior servicio al público, principalmente considerando la importancia que los datos poseen para la región productiva de influencia.

El radar posee sistema doppler y es de doble polarización (DP). Opera en banda C a una frecuencia de 5,64 Ghz y longitud de onda de 5,4 cm [37]. La antena permite un giro en el sentido horizontal (azimut) y puede elevarse en ángulo vertical hasta 45º. Este RM está configurado para completar una serie de giros a 360º que se repite para 12 ángulos de elevación, entre 0,5º de base y 15,1º de tope, en rangos de 120 km, 240 km y 480 km [37], un ejemplo para la primera elevación puede verse en la figura 8(b). La frecuencia de un escaneo completo está programada cada 10 minutos, totalizando 144 adquisiciones diarias normalmente. Cada adquisición se realiza en forma volumétrica, con una unidad de muestreo de $1km^2$ y1°, almacenando *hoy* cada variable, o cómputo derivado, en archivos separados denominados volúmenes. Dentro de las variables que permite recolectar el RM se encuentra el factor de reflectividad (Z), la reflectividad diferencial (ZDR), el coeficiente de correlación polarimétrica (RhoHV), el desplazamiento de fase diferencial (PhiDP), el desplazamiento de fase diferencial específica (KDP), la velocidad radial (V) y la anchura del espectro (W) [16, 38].

En tal sentido debe considerarse que tan solo un RM produce alrededor de 17 GB diarios, lo que arroja aproximadamente un volumen de 6,2 TB por año. Este volumen de datos representa un desafío tanto para su almacenamiento como para su procesamiento y aprovisionamiento en línea a terceros. A la fecha, ante una solicitud de datos en particular a la EEA de INTA Anguil, los mismos deben ser procesados artesanalmente y ad-hoc a partir de los archivos físicos del RM por personal especializado, lo que conlleva en forma implícita a demoras en la respuesta y riesgos propios derivados del procesamiento manual.

En [15, 16] se plantearon los ajustes necesarios sobre la estrategia de procesamiento, a los efectos de incorporar la posibilidad de brindar servicios a terceros y el aprovisionamiento ad-hoc de grándes volúmenes, pero aún no se había definido qué tecnología utilizar para lograr escalabilidad, dinamismo, paralelismo y cómputo distribuido a partir del prototipo original documentado en [13, 32].

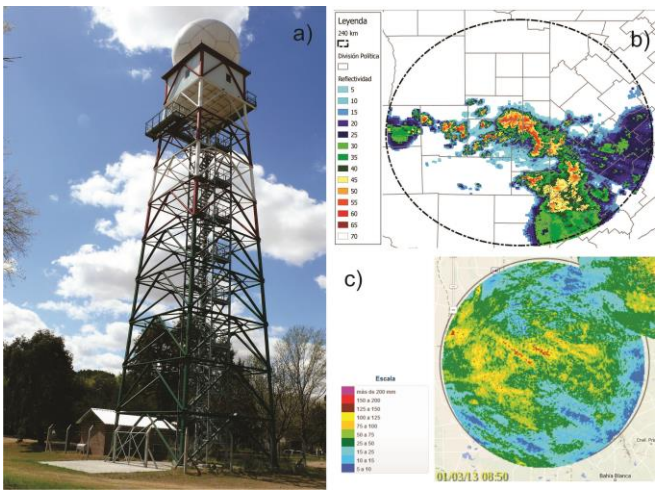

Figura 8. a) Infraestructura del RM instalado en la EEA Anguil, b) Imágen de reflectividad de la primera elevación (0,5°) del 15-01-2011, 23:40hs generado con Software de INTA, c) PAC (Precipitation Accumulation) de Febrero de 2013 generado con Rainbow 5 de Gematronic

De este modo, cada RM es una fuente de datos heterogénea que informará el flujo de medidas mediante C-INCAMI/MIS. Para ello, una pequeña aplicación a instalar en el RM implementa el adaptador de mediciones de APcMM y el Productor de Apache Kafka para un tópico particular, leerá directamente desde el buffer de generación en memoria del RM, informando en tiempo real el flujo. La función de reunión de APcMM actuará como consumidor del flujo C-INCAMI/MIS, efectuando en paralelo: a) La replicación del flujo a terceros mediante suscripción, b) el volcado masivo del flujo en la tabla de medidas específica para el proyecto de M&E dentro del repositorio HBase, y c) La replicación del flujo reunido a la función de análisis y suavización. Éste último, en términos de Apache Storm, sería un *Spout*1 de la topología de procesamiento, cuya estructura queda definida por C-INCAMI/MIS. Luego, tanto el tomador de decisiones, como la función de análisis y suavización, los clasificadores, el consumo y la retroalimentación de la memoria organizacional son *Bolts*1 de la topología Apache Storm, que pueden ser ajustados ante eventuales cambios de los requerimientos de procesamiento, versionándose las topologías e incluso permitiendo ejecutarlas en paralelo y en forma independiente una de otra [4].

Por otro lado, la consulta de medidas y-o de la memoria organizacional almacenada en Apache HBase para los diferentes proyectos de M&E, será provista como servicio por suscripción a partir de Apache Hive. Esto permite, por un lado, el intercambio de datos sin intervención humana, lo que promueve su comunicabilidad y extensibilidad; y por otro, posibilita a los desarrolladores las consultas ad-hoc mediante Hive-QL (acrónimo de *Query Language*) [28], un lenguaje fácil de aprender ya que posee una estructura similar al tradicional SQL (acrónimo de *Structured Query Language*). De

este modo y a partir de tales repositorios, se posibilitarán diferentes análisis de patrones ad-hoc sobre todas las mediciones históricas de los proyectos de la EEA INTA Anguil empleando Apache Mahout [29], lo que permitirá no solo retroalimentar a la APcMM, sino también ajustar las definiciones de los proyectos de M&E como, por ejemplo, lo referido a la emisión de alarmas y/o recomendaciones en vivo.

5.2. Base de Conocimiento para la predicción de pasturas

Esta base de conocimientos almacena un conjunto de casos relacionados al crecimiento del pasto teniendo en cuenta las condiciones climáticas actuales y pasadas, de modo que sirva como base para un sistema de recomendación para la toma de decisiones en un nuevo ciclo de producción de pastos.

La adquisición de conocimiento se realiza a través de CINCAMI/MIS mediante APcMM, donde es posible guardar registro de datos desde diferentes fuentes heterogéneas (Una fuente de datos es el radar meteorológico de la EEA INTA Anguil), así como también, las mediciones manuales realizadas por los productores agrícolas, por ejemplo, la producción diaria estimada de pastos.

Los objetivos de la base de conocimientos es apoyar la productividad, la eficiencia y el crecimiento continuo en este importante sector agrícola de La Pampa; guardando el conocimiento pasado acerca de las condiciones climáticas (como problemas) y la producción de pasto en un determinado periodo (como soluciones). Al proporcionar pronósticos de 60 días sobre el crecimiento de los pastos, sirve como ayuda a los productores de carne para tomar mejores decisiones en la gestión de sus rebaños, producción y costos.

Tabla 1. Ejemplo de modelo de evaluación de similitud para el dominio de predicción de pasturas

Característica	Descripción	Tipo	Peso
Precipitación Acumulada	Lluvia acumulada en los ultimos 10 días	Difference	0.40
Granizo	Indica la ocurrencia de granizo en los ultimos 10 dias (Los valores posibles son *yes* o *no*)	Exact	0.15
Daño por granizo	Indica la ocurrencia de daño por granizo en los ultimos 10 días (Los valores posibles son yes o no)	Exact	0.25
Pronóstico de lluvia	Indica el pronostico de lluvias para los proximos 7 dias	Difference	0.20

Para ilustrar la base de conocimientos, se muestra un modelo simplificado de la estructura del caso de fácil comprensión. Esta base de conocimientos basada en casos caracteriza la situación del problema a través de diversas características incluidas las condiciones meteorológicas actuales previsiones de 7 días, los eventos de lluvia y granizo, y registros del clima pasado (Ver Tabla 1). También los datos de contexto de ubicación y el tiempo se tienen en cuenta. En este ejemplo, el caso se caracteriza por cuatro elementos, a saber: Precipitación adumulada, Granizo, Daño por granizo, y pronóstico de lluvia que están definidas en la ontología de dominio meteorología, que, por razones de espacio no se detalla aquí. Análogamente, la solución se caracteriza por la función de *Producción de pastura*; que indica los kg diarios de materia seca producida en una hectárea.

Para cada rasgo que caracteriza a un caso, debemos establecer su peso y su tipo de función de similitud (véase la Tabla 1). Estas decisiones de diseño son realizadas por un experto teniendo en cuenta que características se consideran más relevantes desde el punto de vista de la similitud para evaluar al final la similitud global de dos casos.

Una vez definida la estructura del caso y su modelo de evaluación de la similitud, cada caso se almacena con todos los valores de rasgos que lo caracterizan y su solución correspondiente. Dos ejemplos de casos se muestran en la Figura 9.

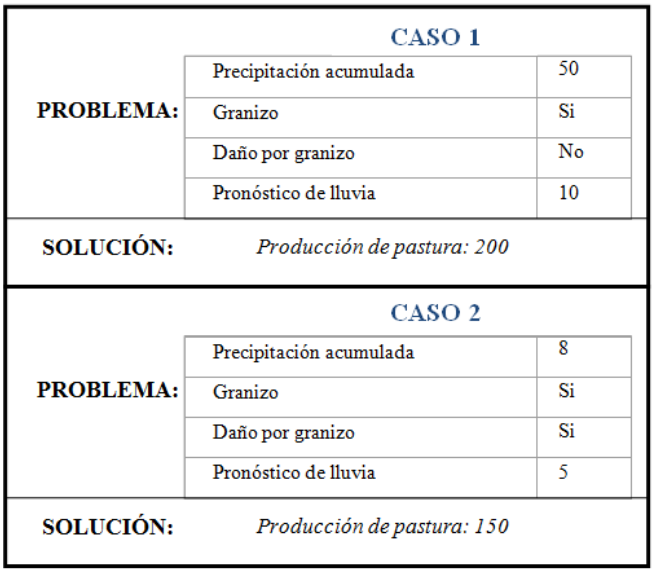

Figura 9. Ejemplo de representacion de dos casos pasados

Una nueva decisión en cuanto al manejo de los rebaños puede ser beneficiada así a partir de la memoria organizacional basada en casos, mediante la recuperación de la información de predicción de pasturas ante condiciones ambientales similares. Supongamos que queremos comprar animales para engorde y necesitamos saber si tendremos suficiente pasto, y adicionalmente contamos en la base de conocimientos con casos tales como los expuestos en la figura 9, entre otros. A los efectos de minimizar riesgos, puede aprovecharse el conocimiento registrado mediante la recuperación y la reutilización de la experiencia pasada con mayor grado de similaridad y así poder respaldar una decisión dada.

La Tabla 2 muestra el cálculo de similitud de cada característica del nuevo caso en comparación con los últimos anteriores, es decir, "Caso 1" y "Caso 2"

Tabla 2. Ejemplo de Evaluación de Similitud entre dos casos previos y uno nuevo

Característica	Caso 1	Caso 2	Nuevo Caso	Similar Caso1/ Nuevo	Similar Caso2/ Nuevo
Precipitación acumulada	50	8	20	0.23	0.83
Granizo	Si	Si	Si	1	1
Daño por granizo	No	Si	No	0	1
Pronóstico de lluvia	10	5	15	0.2	0.1

Por lo tanto, los cálculos de similitud globales dan como resultado:

$$Similaridad(Caso\ 1, Nuevo) = 0.4 * 0.23 + 0.15 * 1 + 0.25 * 0 + 0.2 * 0.2 = 0.282$$

$$Similaridad(Caso\ 2, Nuevo) = 0.4 * 0.83 + 0.15 * 1 + 0.25 * 1 + 0.2 * 0.1 = 0.752$$

Dando como resultado el "Caso 2" como el más similar al nuevo caso, y por lo tanto la predicción de pastos es de 150 kg diarios de materia seca producida. De esta manera, este resultado es utilizado por el tomador de decisiones (Decision Maker en figura 3) a partir de la Memoria Organizacional de por APcMM, para el envío de alarmas o recomendaciones a los interesados que se relacionan con el proyecto de M & E.

6. Trabajos Relacionados

En nuestro trabajo específicamente hemos ilustrado el uso del enfoque de razonamiento basado en casos para desarrollar una memoria organizacional; utilizando la ventaja de tener la arquitectura APcMM, que gestionan grandes volúmenes de datos estructurados, junto con sus metadatos. El objetivo principal es aprovechar el conocimiento que se puede extraer de la memoria organizacional que bajo una estructura clave-valor (es decir, la estructura problema-solución) puede ser procesado eficientemente con las nuevas tecnologías Map Reduce asociada a Big Data; esto permite incorporar más experiencia para recomendar los cursos de acción a las decisiones en el proceso de produccion.

En primer lugar, desde el punto de vista de la memoria organizacional, existen numerosas propuestas en el área de gestión del conocimiento, por ejemplo, los documentados en [39, 40, 41]. La mayoría de ellas capturan y almacenan el conocimiento en repositorios de documentos como manuales, memos y sistemas de archivos de texto, etc., donde rara vez se utilizan estrategias de almacenamiento estructurados o semi-estructurados. Estos enfoques generalmente no emplean potentes mecanismos de procesamiento de conocimiento semántico y automático basado en ontologías, por lo tanto, muy a menudo causan pérdida de tiempo y alta inversión en recursos humanos.

En segundo lugar, desde el punto de la arquitectura de procesamiento de flujo de datos, existen trabajos [42, 1, 6] que enfocan el procesamiento de flujos de datos desde una óptica sintáctica, donde el modelo de datos del flujo se basa en una estructura clave-valor e incorporan técnicas para la gestión adaptativa de tasas de arribo, para poder abordar el tratamiento de volúmenes de datos explosivos en los flujos [41]. Nuestra arquitectura incorpora la capacidad de introducir metadatos basados en un marco formal de M&E, que guian la organización de las medidas (por ejemplo, mediante instantáneas en memoria y el último estado conocido de la entidad bajo análisis), facilitando análisis consistentes y comparables desde el punto de vista estadístico, con la posibilidad de disparar alarmas basada en la interpretación de los criterios de decisión de los indicadores que se obtienen a partir de los datos. No obstante, Lee y otros [43] plantean una interesante limitación de Apache Storm para abordar los flujos de datos explosivos, que será motivo de estudio y verificación empírica en APcMM. Adicionalmente, nuestra propuesta cuenta con los procesos formalizados mediante SPEM, lo que promueve una especificación bien establecida, comunicable y extensible.

7. Conclusiones y Trabajo Futuro

La gestión del conocimiento organizacional representa un activo clave como apoyo a un más eficaz proceso de toma de decisiones por parte de diferentes grupos de interés. En este sentido, tener una memoria organizacional basada en casos que soporta la estructuración, y el procesamiento automático del conocimiento de la organización es una decisión primaria para lograr su gestión eficaz. En este trabajo ha planteado una memoria organizacional basada en casos para la arquitectura de procesamiento de flujo de datos basada en metadatos de medición. La representación del conocimiento a través de casos facilita la reutilización de los conocimientos adquiridos en los problemas del pasado que se aplicará a un nuevo problema en situaciones similares.

Utilizando dicha memoria organizacional se desarrolló un sistema de recomendación que consiste en la predicción de pasturas para la producción agrícola. De esta manera, el sistema de recomendacion puede beneficiarse de la potencia de procesamiento de la arquitectura APcMM, la cual sustenta su funcionamiento en tecnologías Big Data distribuida. También se ha sintetizado la arquitectura APcMM, la importancia de los metadatos como guías del procesamiento en tiempo real. Adicionalmente, la posibilidad de informar datos y metadatos en forma conjuntamenta promueve la interoperabilidad de los datos por cuanto no solo pueden ser intercambiados, sino comprendidos e interpretados a partir de la definición del proyecto de M&E en base a C-INCAMI.

Así, los flujos de datos obtenidos a partir de fuentes heterogéneas, los datos procesados y-o provistos en tiempo real por suscripción, o bien aquellos almacenados en grandes repositorios, se estructuran bajo el esquema C-INCAMI/MIS, con el objetivo de mejorar la interoperabilidad del sistema. Adicionalmente y a los efectos de la interoperabilidad, debe mencionarse que la arquitectura tiene formalizado sus procesos en base a SPEM, lo que fomenta su comunicabilidad y extensibilidad

Por último, hemos ilustrado estos modelos y enfoque con un caso práctico: un sistema de predicción de pasturas, utilizando los datos del radar meteorológico de la EEA INTA

Anguil. Esta base de casos almacena un conjunto de conocimientos relacionados con el crecimiento de los pastos en base a una serie de datos, incluyendo las condiciones actuales y pronósticos meteorológicos, eventos de lluvia y los registros climáticos del pasado, procesados por la arquitectura APcMM y considerando al radar como una de las fuentes de datos.

Como trabajo a futuro, se avanzará en la implementación de la APcMM en relación a la recolección de datos distribuida mediante C-INCAMI/MIS y los adaptadores de mediciones a través de Apache Kafka.

8. Agradecimientos

Esta investigación está soportada por los proyectos PICTO 2011-0277 de la Agencia de Ciencia y Tecnología de la nación y el proyecto 09/F068 de la Facultad de Ingenieria. UNLPam.

9. Referencias

[1] Chakravarthy, S. and Jiang, Q., Stream Data Processing: A Quality of Service Perspective. Springer, 2009.

[2] Guller, M., Big Data Analytics with Spark. A Practitioner's Guide to Using Spark for Large Scale Data Analysis. New York, USA: Apress, 2015.

[3] Diván, M., Olsina, L., and Gordillo, S., "Strategy for Data Stream Processing Based on Measurement Metadata: An Outpatient Monitoring Scenario," Journal of Software Engineering and Applications, vol. 4, no. 12, pp. 653-665, December 2011.

[4] Jain, A. and Nalya, A., Learning Storm. Create real-time stream processing applications with Apache Storm. Birmingham, United Kingdom: Packt Publishing Ltd., 2014.

[5] Frampton, M., Mastering Apache Spark. Birmingham, United Kingdom: Packt Publishing Ltd., 2015.

[6] Cugola, G and Margara, A "Processing flows of information: From data stream to complex event processing," Journal of ACM Computing Surveys, vol. 44, no. 3, p. Article No. 15, June 2012.

[7] Bockermann, C. and Blom, H. "Processing Data Streams with The RapidMiner Streams Plugin," Technical University of Dortmund, Dortmund, Germany, Report 2012.

[8] Diván, M., Martín, M., and Olsina, L., "Hacia la Retroalimentación del Procesamiento de Flujos de Datos Sustentado en Memoria Organizacional," in Primer Congreso Nacional de Ingeniería Informática / Sistemas de Información, Córdoba, Argentina, 2013, pp. 79-90.

[9] Diván, M and Olsina, L., "Process View for a Data Stream Processing Strategy based on Measurement Metadata," Electronic Journal of SADIO, vol. 13, no. 1, pp. 16-31, June 2014.

[10] Molina, H. and Olsina, L., "Towards the Support of Contextual Information to a Measurement and Evaluation Framework," in QUATIC, Lisboa, Portugal, 2007, pp. 154–163.

[11] Olsina, L, Papa, F., and Molina, H., "How to Measure and Evaluate Web Applications in a Consistent Way," in Ch. 13 in Web Engineering.: Springer, 2007, pp. 385–420.

[12] Abajo Martínez, N., "ANN quality diagnostic models for packaging manufacturing: an industrial data mining case study," in ACM Special Interest Group on Knowledge Discovery and Data Mining (SIGKDD), Seattle (USA), 2004, pp. 799-804.

[13] Diván, M., "Enfoque Integrado de Procesamiento de Flujos de Datos centrado en Metadatos de Mediciones," UNLP, La Plata, PhD Thesis 2011.

[14] Diván, M. and Olsina, L., "Process View for a Data Stream Processing Strategy based on Measurement Metadata," Electronic Journal of Informatics and Operations Research, vol. 13, no. 1, pp. 16-34, June 2014.

[15] Diván, M. and Martín, M., "Estrategia de Procesamiento de Flujos de Datos Sustentada en Big Data y Memoria Organizacional," in Tercer Congreso Nacional de Ingeniería Informática / Sistemas de Información, Buenos Aires, 2015.

[16] Diván, M, Bellini Saibene, Y., Martín, M, Belmonte, L., Lafuente, G. and Caldera, J., "Towards a Data Processing Architecture for the Weather Radar of the INTA Anguil," in International Workshop on Data Mining with Industrial Applications, Asunción, Paraguay, 2015.

[17] Marrocco, C, Duin, R., and Tortorella, F. "Maximizing the area under the ROC curve by pairwise feature combination," ACM Pattern Recognition, pp. 1961-1974, 2008.

[18] SPEM, "Software Process Engineering Meta-Model Specification," Object Management Group (OMG), Ver.2.0, 2008.

[19] Apache Software Foundation. (2016, Abril) Apache Kafka. [Online]. http://kafka.apache.org/

[20] Garg, N., Apache Kafka. Set up Apache Kafka clusters and develop custom message producers and consumers using practical, hands-on examples. Birmingham, United Kingdom: Packt Publishing Ltd., 2013.

[21] Apache Software Foundation. (2016, Abril) Apache Storm. [Online]. http://storm.apache.org/index.html

[22] Apache Software Foundation. (2016, Abril) Apache HBase. [Online]. http://hbase.apache.org/

[23] Spaggiari, J and O'Dell, K, Architecting HBase Applications (Early Release). New York, USA: O'Reilly Media, 2015.

[24] R Core Team, R: A Language and Environment for Statistical Computing. Vienna, Austria: The R Foundation for Statistical Computing, 2016.

[25] Da Silva, M and Tavares, H, Redis Essentials. Birmingham: Packt Publishing, 2015.

[26] Bifet, A., Holmes, G., Kirkby, R., and Pfahringer, B., "MOA: Massive Online Analysis," Journal of Machine Learning Research, vol. XI, pp. 1601-1604, 2010.

[27] Apache Software Foundation. (2016, Abril) Apache Hadoop. [Online]. http://hadoop.apache.org/

[28] Rutherglen, J, Wampler, D, and Capriolo, E, Programming Hive. California: O'Reilly Media Inc., 2012.

[29] Gupta, A., Learning Apache Mahout Classification. Birmingham: Packt Publishing, 2015.

[30] Holmes, A, Hadoop in Practice, 2nd ed. New York, USA: Manning, 2015.

[31] Martín, M. and Olsina, L. "Added Value of Ontologies for Modeling an Organizational Memory," in Building Organizational Memories: Will You Know What You Knew?, John Girard, Ed. USA: IGI Global, 2009, ch. 10, pp. 127-147.

[32] Martín, M. "Memoria Organizacional Basada en Ontologías y Casos para un Sistema de Recomendación en Aseguramiento de la Calidad," Facultad de Informática, Unviersidad Nacional de La Plata, La Plata, PhD Thesis 2011.

[33] Kolodner, J. Case-based Reasoning.: Morgan Kaufmann, 1993.

[34] Chen, H and Wu, Z. "On Case-Based Knowledge Sharing in Semantic Web," in XV International Conference on Tools with Artificial Intelligence, California, 2003, pp. 200-207.

[35] Aamodt, A. and Plaza, E., "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches," Artificial Intelligence Communications, vol. 7, no. 1, pp. 39-59, 1994.

[36] Coyle, L., Doyle, D. and Cunningham, P. "Similarity for Case-Based Reasoning," Trinity College, Dublin, Technical Report TCD-DS-2004-25, 2004.
[37] Hartmann, T., Tamburrino, M. and Bareilles, S. "Preliminar Analysis of data obtained from the INTA radar network for the study of the precipitations in the pampeana region (in spanish)," in 39º Jornadas Argentinas de Informáticas - 2º Congreso Argentino de Agroinformática, Buenos Aires, Septiembre 2010, p. 826.
[38] Gematronik, Instruction Manual. Rainbow 5. Neuss, Germany: Gematronik GmbH, 2007.
[39] Conklin, J "Designing Organizational Memory: Preserving Intelectual Assets in a Knowledge Economy," Group Decision Support Systems, http://www.gdss.com/DOM.htm 1996.
[40] Lindstaedt, S. Strohmaier, M., Rollett, H., Hrastnik, J., Bruhnsen, K., Droschl, G. and Gerold, M., "KMap: Providing Orientation for Practitioners when Introducing Knowledge Management," in Practical Aspects of Knowledge Management: 4th International Conference, PAKM 2002 Vienna, Austria, December 2--3, 2002 Proceedings, D Karagiannis and U Reimer, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, vol. 2569, pp. 2-13.
[41] Karl-Heinz, W., "A Case Based Reasoning Approach for Answer Reranking in Question Answering," CoRR, vol. abs/1503.02917, 2015. [Online]. http://arxiv.org/abs/1503.02917
[42] Botan, I., Derakhshan, R., Dindar, N., Haas, L., Miller, R. and Tatbuk, N., "SECRET: a model for analysis of the execution semantics of stream processing systems," In proc. of VLDB Endowment, vol. 3, no. 1-2, pp. 232-243, September 2010.
[43] Lee, M., Lee, M., Hur, S., and Kim, I. "Load Adaptive and Fault Tolerant Distributed Stream Processing System for Explosive Stream Data," Transactions on Advanced Communications Technology, vol. 5, no. 1, pp. 745-751, 2016.