**Universidad del Cauca**

**Facultad de Ingeniería Electrónica y Telecomunicaciones**

**Programas de Maestría y Doctorado en Ingeniería Telemática**

**Seminario de Investigación**

**FRAMEWORK FOR DATA QUALITY IN KNOWLEDGE DISCOVERY TASKS (FDQ-KDT)**

**Relator: Mag. Camilo Corrales, estudiante de Doctorado**

**Co-relator: Ph.D. Juan Carlos Corrales**

**Protocolante: Mag. Mario F. Solarte S., estudiante de Doctorado**

**Fecha:** 6 de marzo de 2015

**Hora inicio:** 10.15 a.m.

**Hora fin:** 11:30 m

**Lugar:** Salón de posgrado, FIET, Universidad del Cauca, Popayán

**Asistentes:**

Ph.D. Juan Carlos Corrales, coordinador del seminario

Estudiantes de Maestría y Doctorado en Ingeniería Telemática

Estudiantes de pregrado de la FIET

Ph.D. Oscar Mauricio Caicedo

**Orden del día:**

1. Presentación a cargo del relator
2. Intervención del co-relator
3. Discusión
4. Cierre

**Desarrollo**

1. **Presentación a cargo del relator**

Empieza Camilo Corrales por indicar la agenda de su presentación, que consiste en:

a.- Contexto y conceptos importantes

b.- Escenario de motivación

c.- Trabajos relacionados

d.- Objetivos

1. **Contexto y conceptos importantes**

Empieza por brindar algunas definiciones necesarias para entender el desarrollo de la propuesta.

Descubrimiento de conocimiento (DC): proceso no trivial, en el cual se debe tener un conjunto de datos (base fundamental) y encontrar patrones sobre ese conjunto de datos (Usama et al,2006). Existen muchas tareas en CD. Me voy a centrar en clasificación y agrupamientos.

Hay muchas técnicas para hacerlo, yo me voy a centrar en las técnicas: Clasificación y Agrupamiento.

Cuanco queremos hacer una tarea de clasificación: predicción, clasificación o diagnóstico, Ese está compuesto de atributos y una columna adicional que es el objetivo (lo que se quiere predecir o clasificar). El Agrupamiento no tiene en cuenta esa última columna.

En el dataset, las filas son llamas instancias: variaciones de atributos y de la clase

Framework: representación de componentes principales para abordar un problema de interes, el framework debe representar la relaciones entre los componentes de ese problema *Georgina et al, 2011*. Que sirva de plataforma de entendimiento entre los usuarios.

Framework calidad de datos: herramienta para la evaluación de la calidad de datos en una organización Wang et al, 1996. Proporciona un esquema para organizar y resolver la calidad dee los datos para una gestión proactiva (*Eppler et al, 2000)*. Evaluar donde hay áreas carencia en la calidad de los datos, reducir los problemas dentro de la organización. *(Kerr et al, 2004)*

**b.- Escenario de motivación**

En el mundo cotidiano hay muchas tareas de agrupación o clasfficación: predicción de clima, categorizar correos electrónicos. Como insumo necesita un dataset, ingreso a un algoritmo genera una agrupación, o modelos.

¿Qué pasa con los datos que tiene baja calidad?

Esas agrupaciones, modelos y funciones de los algoritmos son imprecisos y pueden tener ruidos.

Problemas: en la construcción de datos puede haber errores por entradas manuales. Si son sensores, se pueden descalibrar. La transferencia de datos de un lugar a otros puede hacerse con pérdidas.

Las técnicas carecen de técnicas para evaluar la calidad de datos.

En ese sentido voy a abordar las metodologías más utilizadas para poder levar un proceso formal en tareas de DC.

CRISP-DM: conjunto de fase: mantenimiento, negocio, datos, prepación, modelaje, evaluación y despliegue también hay subtareas (*Chapman et al, 2000)*; aquí están resaltadas en rojo las que tienen que ver con la calidad de los datos .

SEMMA: a su vez tiene cinco fases: muestreo, exploración, modificación, modelado (con los algoritmos), y finalmente se hace una evaluación del modelo *(Olson et al, 2008)*. Se resaltan las actividades que tienen que ver con la calidad de los datos.

Ahora está pegando el tema denominado "Ciencia de los Datos". La Universidad John Wopkins ha generado un flujo de trabajo para hacer Ciencia de los Datos que se ofrece como una especiación en línea; se utilizan técnicas de diversos dominios: minería de datos, inteligencia artificial, estadística (*O’Neil and R. Schutt, , 2013)*. Parte del hecho que se tiene un conjunto de datos, se limpian y se analizan para verificar cómo funcionan los datos.

Brecha 1, a nivel de marco teórico: ¿Las metodologías abordan la calidad de los datos, muchas de ellas dejan por encima o muy suelo... no aclaran los problemas, hay que conocer los datos? *(Pacheco et al, 2014.)*

**c.- Trabajos relacionados**

Taxonomía del 2013: problemas que se pueden presentar en un conjunto de datos. Hecha y enfocada para Ingeniería de Software Empírico: Precisión, Relevancia, Proveniencia*(Bosu and MacDonell,2013)*.

En este sentido, se descarta la proveniencia porque son datos del modelo del negocio, a los cuales generalmente no se puede acceder.

**Precisión**

- **Valores atípicos**: observaciones fuera de la distribución normal de los datos.

- **Ruido**: información irrelevante de los datos.

- **Inconsistencias**: falta de armonía en los datos. Instancias contradictorias.

- **Incompletitud**: faltan valores en el conjunto de datos.

- **Redundancia**: instancias duplicadas

**Relevancia**

- **Cantidad de datos**: si hay un dataser pequeño, los resultados no van a ser buenos para un algoritmo de aprendizaje de máquina. Dataset imbalanceados

- **Heterogeneidad**: problemas que ocurren cuando se intenta fusionar datos de distintas fuentes.

- **Puntualidad**: el más complejo de encontrar en la literatura. Se tiene un conjunto de datos que contienen datos correspondientes a condiciones atípicas.

De cada uno de ellos indica las fuentes bibliográficas consultadas, los artículos encontrados, los descartados, la clasificación y tendencias temáticas y temporales encontrados a partir del mapeo sistemático.

Brecha 2: no se ha encontrado una herramienta que integre todos estos problemas

Brecha 3: no en todos los casos se sabe cuál es el mejor algoritmo para afrontar un problema particular.

¿Cómo mejorar la calidad de los datos en tareas de descubrimiento de conocimiento a través de algoritmos de inteligencia artificial?

**d.- Objetivos:**

**General:**

Definir un framework conceptual que identifique esos problemas en las tareas de DC.

**Específicos:**

Construir un mecanismo que reúna esas técnicas de inteligencia artificial para resolver los problemas de identificación de datos identificados en el framework

Establecer estrategias que aconsejen cuál es el algoritmo adecuado a utilizar para resolver un problema determinado en calidad de datos

Desarrollar y evaluar experimentalmente un prototipo que pruebe todo el framework.

Finalmente, quiero comentar que hace dos días se notificó la aceptación de un artículo en un evento en Chicago, que incluye la propuesta de Doctorado, Cuarta conferencia sobre Bases de Datos y Minería de Datos, se va a publicar en la Revista Journal of Computer, Q2 SIMAGO, A2 en Publindex.

*DC. Corrales, A. Ledezma, JC. Corrales, “Framework for data quality in knowledge discovery tasks (FDQ-KDT): a proposal”, in 4th International Conference on Database and Data Mining , 2015.*

1. **Intervención del co-relator**

Trabajo que se está llevando a cabo en colaboración con U3CM con el profesor Agapito Ledezma, se firmó convenio marco y se está formulando el convenio específico para con el mismo trabajo poder sustentar en ambas partes.

Señalo la elaboración el mapeo sistemático, y en la próxima vuelta se hablará más acerca de la herramienta a construir.

Tratamos de ir rápido en el tema de la revisión para posicionar la propuesta de alguna manera, el paper fue aceptado; se han ido desprendiendo trabajos de pregrado (calidad del agua) y de maestría (algoritmos incrementales). Ha rendido estos seis meses.

Hay que revisar el objetivo tres, ¿por qué no hacer algo como un sistema de recomendación, aunque sin ser un sistema de recomendación, plantee recomendaciones a los usuarios? La alternativa parece ser usar planeadores. El problema es que usar dataset liga mucho al dominio de aplicación, y precisamente se quiere estar lo más independiente posible del dominio de aplicación.

1. **Discusión**

**Juan PabloRuiz**: no me queda claro la selección del set de datos de pruebas.

**Camilo Corrales**: nos importa tener un framework. MI tutor de España quiere usar un dominio de Netflix de un millón de datos. La idea es wque el framework esté por encima del dominio de aplicación.

**JPR**: en inteligencia artificial tienes algoritmos supervisados... ¿cómo se haría esa supervisión para que los algoritmos aprendan esas clasificaciones?

**CC**: a partir de un sistema experto puedes plantear un conjunto de reglas que generen datos sintéticos, con ellos puedes hacer agrupamientos y de ahí clasificación para detectar tipos de problemas en un dominio particular (valores atípicos). Los valores atípicos se manejan con algoritmos de agrupamientos.

**JPR:** ¿Habrá algún experto o ser humano pendiente de que esté aprendiendo bien?

**CC**: la idea es que sea asistido. Se sabe que es imposible tener un proceso automatizado completamente para calidad de datos.

**Juan Carlos Corrale**s: nosotros queremos aplicar al dominio de la agricultura de precisión, pero mientas se consigue la información, se pueden usar otros tipos de datos con dataset publicados, por ejemplo con Netflix.

**CC**: lo interesante de este trabajo es que surgió a partir del trabajo de maestría, yo estaba haciendo estimaciones con 160 instancias cuando para clasificación o clusterin se necesitan 16000 instancias para poder hacer simulaciones.

**Gustavo Uribe**: de todas maneras debes hacer alguna caracterización o clasificación de los dominios. Hay una explosión de algoritmos y la única explicación es que haya una dependencia de los dominios.

**CC**: es cierto, en la revisión es ir acotando esa parte, por lo menos se está acotando. Lo importante es caracterizar los dominios, pero al final el dominio da un comportamiento al conjunto de datos.

**JCC**: Ahí está la discusión con Agapito... si hay que tener una Ontología para tener una referencia, a los quince días que no... que mejor con Sistema Basados en Casos, luego llegaron los artículos sobre Planeadores. Hay para revisar un poco e identificar el foco.

**GU**: No me queda clara la diferencia entre valores atípicos y ruido.

**CC**: los valores atípicos pueden ser valores correctos sólo que por fuera de los datos normales. Aunque eso depende de los autores... algunos incluyen los outlayers como valores atípicos.

**GU**: el ruido, aunque no se espera, puede tener otras características identificables en los datos recolectados.

**Gabriel Chanchí**: tengo una duda respecto al objetivo 2. ¿No saben si es demasiado ambicioso trabajar con todos los algoritmos que pueden salir? Podrían puntualizar en tema del contexto... si son genéricos, no encuentro la forma cómo podría trabajar con algoritmos genéricos si van involucrados en la solución.

**CC**: eso se ha discutido... Agapito Ledezma nos pregunta ¿por qué deben hacer un anteproyecto si los objetivos pueden cambiar a lo largo de la investigación? Por eso se dejan genéricos. Por debajo del contexto, los comportamientos matemáticos de los problemas de los dataset son los mismos que pueden aparecer en casi cualquier contexto.

**Oscar Mauricio Caicedo**: creo que el objetivo dos está acotado implícitamente por el objetivo 3, donde vas a hacer la selección de los algoritmos adecuados para el análisis de la calidad, va a hacer una colección de algoritmos y unas estrategias para la selección de las estrategias que apuntarán a cuál es el mejor dependiendo del contexto. Es más interesante que se deje genérico, porque precisamente eso es lo que no se ha hecho... ese es el reto del doctorado.

**JCC**: por ejemplo, si lo hubiéramos dejado sólo aplicado a la agricultura, no creo que se hubiera aceptado el paper en el Evento de Chicago, hubiéramos tenido que buscar otro más aplicado al sector. De la solución genérica se puede pasar a otro dominio

**OMC**: Pensando en el marketing de los papers es más atractivo que sea una solución genérica con un ejemplo de aplicación.

Me gustó bastante tu presentación. El trabajo de revisión bibliográfica parece bien hecho, tengo la sugerencia para intentar un paper de revisión, número alto de papers y análisis. Tomaste unas definiciones, las articulaste, ese trabajo se puede pasar en una publicación Surveys and Tutorials de impacto 8, hay que filtrar las referencias por impacto de las conferencias o sitios donde se publicaron... o sea, limpiar los datos de las referencias que has utilizado.

Son muy difícil de pasar, pero con todo el material que tienes, sería interesante intentar. Pueden ser paper con 200 referencias.

Sobre la pregunta de investigación... ¿en qué términos defines la calidad de los datos?

**CC**: mirar cómo se comportan los datos con un algoritmo de clasificación, por ejemplo, limpiarlos y ver cómo se comportan después.

**OMC**: sugiero cambiar la palabra por otra... también te daría claridad sobre cómo evaluar el framework. ¿Cómo vas a medir la efectividad del framework? En mi caso, era la eficiencia del consumo del ancho de banda, tiempo de respuesta. Sólo para que lo tengan en cuenta... por eso no me gustan los anteproyectos y uno debe tener flexibilidad, pero si hay que tener en cuenta la forma en la cual se va a hacer la evaluación.

¿Cuál sería tu hipótesis?

**CC**: mejorar la calidad de los datos.

**OMC**: ese ejercicio es bien interesante... ¿qué hipótesis se desprende de la pregunta de investigación?

1. **Cierre**

El coordinador del seminario da fin a la discusión y termina la sesión, invitando a asistir a la sesión de la próxima semana y agradeciendo la asistencia del público.