**Universidad del Cauca**

**Instituto de postgrados en Ingeniería Electrónica y Telecomunicaciones**

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

**Seminario de Investigación**

***Titulo de la relatoría:***

**Framework for data quality in knowledge discovery tasks (FDQ-KDT)**

**Relator: Msc. Camilo Corrales estudiante de Doctorado**

**Co-relator: PhD. Juan Carlos Corrales**

**Protocolante: Msc. Maritza Mera, estudiante de Doctorado**

**Fecha:** 13 de noviembre de 2015

**Hora de Inicio:** 10:15 a.m.

**Hora de Finalización:** 11:15 a.m.

**Lugar:** Universidad del Cauca, Popayán

**Asistentes:**

PhD. Oscar Mauricio Caicedo (Coordinador encargado del seminario)

PhD. Juan Carlos Corrales (Co-relator)

MSc. Camilo Corrales (Relator)

Estudiantes de Maestría y Doctorado en Telemática (U. del Cauca)

Estudiantes de Pregrado de la FIET

**Orden del día:**

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

**Desarrollo:**

1. **Presentación a cargo del relator:**

El MSc. Camilo Corrales inicia la presentación de su propuesta de investigación de doctorado, mencionando los temas que abarcará durante la misma, los temas definidos son los siguientes: contexto, motivación, avance de la investigación, y las contribuciones de investigación. A continuación se explica en detalle cada tema expuesto durante la presentación.

**Contexto en Internet de las Cosas**

El Msc Corrales hace una contextualización presentando algunas definiciones que soportan el desarrollo de su investigación. La primera definición que presenta es sobre Descubrimiento de Conocimiento, con base en lo definido por Fayyad et al en 1996, se define *Descubrimiento de Conocimiento como un proceso no trivial de identificar patrones comprensibles, validos, útiles y actualizados desde grandes colecciones de datos.* A continuación, el Msc Corrales destaca que una de las tareas más importantes en el Descubrimiento de la información es es la extracción de conocimiento de los datos.

Las tareas que definen el proceso de Descubrimiento de Conocimiento son

* Reglas de Asociación
* Clasificación
* Regresión
* Análisis de cluster

Una vez explicado el proceso de descubrimiento de Conocimiento se definen los conceptos de Framework y Framework de Calidad de los Datos.

Framework

Según Georgina (2011), es una representación de los principales componentes de un sistema o tema de interés, mostrando sus interrelaciones. Esto sirve para desarrollar un entendimiento común de temas que deberían ser incluidos en una evaluación.

Framework de Calidad de los Datos

Según Wang et al (1996), un Framework de calidad de datos es una herramienta para la evaluación de calidad de datos dentro de una organización.

Teniendo en cuenta lo anterior, el Msc presenta la taxonomía de los retos de la calidad de los datos dentro de la Ingeniería del Software Empírica. Esta taxonimía puede observarse en la siguiente gráfica:

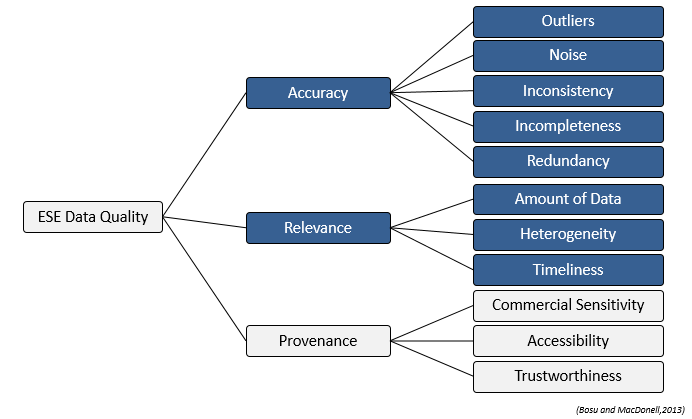


Figura 1. A Taxonomy of Data Quality Challenges in Empirical Software Engineering (ESE)

Posteriormente, el Msc presenta los problemas de Calidad de los Datos en Ingeniería del Software Empírica. Los cuales son los siguientes:

* Exactitud
  + Aislados
  + Ruido
  + Inconsistencia
  + Incompletitud
  + Redundancia
* Relevancia
  + Cantidad de datos
  + Heterogeneidad
  + Oportunidad (tiempo)

**Motivación**

Luego de la contextualización presentada, el Msc. Corrales expone el escenario de motivación de su trabajo. Destacando principalmente que la mala calidad de los datos tienen un impacto en la calidad de los resultados de análisis en las tareas de descubrimiento de conocimiento y en consecuencia, esto impactará sobre las decisiones realizadas y soportadas sobre esos resultados.

**Progreso de investigación**

El Msc Corrales presenta los objetivos de su trabajo y los avances realizados en cada uno de estos. Los objetivos se mencionan a continuación:

**Objetivo General**

Desarrollar un Framework para calidad de datos en tareas de descubrimiento de conocimeinto a través de algoritmos de inteligencia artificial

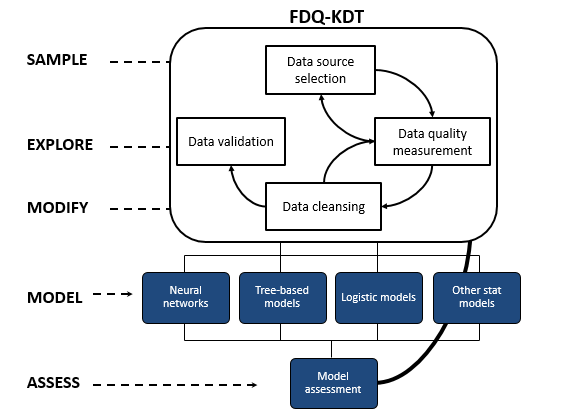
**Objetivos específicos**

* Definir un framework conceptual que analice los problemas de calidad de los datos en tareas de descubrimiento de conocimiento.
* Construir un mecanismo que reúna algoritmos de inteligencia artificial para solucionar los problemas de la calidad de los datos identificados para el framework
* Establecer estrategias que asesoren al usuario los algoritmos de inteligencia artificial adecuados para resolver el problema de calidad de los datos.
* Desarrollar y evaluar experimentalmente un prototipo que pruebe los mecanismos y estrategias del framework en las tareas de descubrimiento de conocimiento.

A partir de los objetivos presentados, el Msc Corrales presenta los tres trabajos en los que ha basado su investigación. Los cuales son:

* **Quality awareness for managing and mining data,** Laure Berti-Equille. University of Rennes. June 2007.
* **Desarrollo de software dirigido por modelos para facilitar a usuarios inexpertos la aplicación de técnicas de minería de datos.** Roberto Espinosa Oliva. University of Alicante. November 2014.
* **Dataset analysis for classifier ensemble enhancement.** Emanuele Tamponi. University of Cagliari. April 2015.

Teniendo en cuenta los trabajos estudiados, el Msc Corrales propone el framework de calidad de los datos considerando las etapas de CRISP-DM



**CONTRIBUCIONES**

1. David Camilo Corrales, Agapito Ledezma and Juan Carlos Corrales, A conceptual framework for data quality in knowledge discovery tasks (FDQ-KDT): a proposal, Journal of Computers, November 2015.
2. David Camilo Corrales, Apolinar Figueroa, Agapito Ledezma and Juan Carlos Corrales, An empirical multi-classifier for coffee rust detection in Colombian Crops, Computational Science and Its Applications – ICCSA 2015, Banff, AB, Canada.
3. David Camilo Corrales, Apolinar Figueroa and Juan Carlos Corrales, Towards detecting crop diseases and pest by supervised learning, Revista Ingeniería y Universidad – Universidad Javeriana, June 2015.
4. Edwin Castillo, Fernando Gonzales, Iván López, Apolinar Figueroa , Miller Guzmán, David Camilo Corrales, and Juan Carlos Corrales, Water quality warnings based on cluster analysis in Colombian river basins, Revista Sistemas & Telemática – Universidad ICESI, June 2015.
5. David Camilo Corrales, Apolinar Figueroa, Agapito Ledezma and Juan Carlos Corrales, Two-level classifier ensembles for coffee rust estimation in Colombian crops, International Journal of Agricultural and Environment Information Systems , January 2016.
6. David Camilo Corrales, Agapito Ledezma and Juan Carlos Corrales, A systematic review of data quality issues in knowledge discovery tasks, Revista Ingenierías Universidad de Medellín, January 2016.

**DISCUSIÓN**

El Dr Juan Carlos Corrales inicia la co-relatoria, y justifica la las razones de emplear los tres trabajos que el Msc. Corrales presentó la presentación. Destaca la importancia del Dataset con el que se van a realizar las pruebas. Explica que posterior a la realización de las pruebas se va a empezar a detallar en artículos el funcionamiento de cada componente definido en el modelo.

A continuación se inicia la discusión.

El Msc Diego Durán pregunta sobre la importancia del dominio del origen de los datos dentro del framework

El Msc Corrales responde que el concepto es plantear un framework genérico. Sin embargo, que se debe seleccionar un conjunto de datos de prueba. El objetivo es que el framework no esté ligado a un dominio de aplicación.

El Dr Corrales afirma que se debe probar el framework en varios dominios de aplicación. Adicionalmente, las tesis de maestría y pregrado están ayudando a alimentar un dataset sobre datos de agricultura. Sin embargo, las publicaciones deberán tener varios dominios de aplicación.

El Msc Diego Durán pregunta sobre la relevancia de las bases de datos en el framework.

El Msc Corrales responde que esto influye en la manera como se representa la información en el framework. Aquí aplican todos los conceptos del modelo relacional. De esta manera, el framework está enfocado hacia el problema semántico y no sintáctico de las bases de datos.

La Msc Alexandra Gómez pregunta sobre cómo serán las pruebas del framework

El Msc Corrales responde que aún se encuentra en revisión está parte.

El Dr. Corrales dice que el Msc Corrales tiene como tarea realizar un plan de pruebas para cada uno de los componentes que describe su modelo.

**REFERENCIAS**

- Pérez-Ariza, C.B., A.E. Nicholson, and M.J. Flores, *Prediction of Coffee Rust Disease Using Bayesian Networks*, in *The Sixth European Workshop on Probabilistic Graphical Models*, M.G.-O. Andrés Cano, Thomas D. Nielsen, Editor 2012, DECSAI, University of Granada: Granada (Spain).

- Cintra, M.E., et al. *The use of fuzzy decision trees for coffee rust warning in Brazilian crops*. in *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*. 2011.

- Luaces, O., et al., *Using nondeterministic learners to alert on coffee rust disease.* Expert Systems with Applications, 2011. 38(11): p. 14276-14283.

- Luaces, O., et al., *Viability of an alarm predictor for coffee rust disease using interval regression*, in *Proceedings of the 23rd international conference on Industrial engineering and other applications of applied intelligent systems - Volume Part II*2010, Springer-Verlag: Cordoba, Spain. p. 337-346.

- Kaundal, R., A. Kapoor, and G. Raghava, *Machine learning techniques in disease forecasting: a case study on rice blast prediction.* BMC Bioinformatics, 2006. 7: p. 485.

- Jain, R., S. Minz, and Ramasubramanian, *Machine Learning for Forewarning Crop Diseases.* Journal of the Indian Society of Agricultural Statistics 2009. 63: p. 97-107.

- Korada, N.K., N.S.P. Kumar, and Y.V.N.H. Deekshitulu, *Implementation of Naive Bayesian Classifier and Ada-Boost Algorithm Using Maize Expert System.* International Journal of Information Sciences and Techniques (IJIST), 2012. 2.

- Paul, P.A. and G.M. Munkvold, *A Model-Based Approach to Preplanting Risk Assessment for Gray Leaf Spot of Maize.* The American Phytopathological Society: Ecology and Epidemiology, 2004. P-2004-1011-04R.

- Guerrero, J.M., et al., *Support Vector Machines for crop/weeds identification in maize fields.* Expert Systems with Applications, 2012. 39(12): p. 11149-11155.

- Meira, C.A.A., L.H.A. Rodrigues, and S.A.d. Moraes, *Modelos de alerta para o controle da ferrugem-do-cafeeiro em lavouras com alta carga pendente.* Pesquisa Agropecuária Brasileira, 2009. 44: p. 233-242.

- Meira, C.A.A. and L.H.A. Rodrigues, *Árvore de decisão na análise de epidemias da ferrugem do cafeeiro.* VI Simpósio de Pesquisa dos Cafés do Brasil, 2009.

- Meira, C., L. Rodrigues, and S. Moraes, *Análise da epidemia da ferrugem do cafeeiro com árvore de decisão.* Tropical Plant Pathology, 2008. 33(2): p. 114-124.

- Liaw, S.T., et al., *Towards an ontology for data quality in integrated chronic disease management: A realist review of the literature.* International Journal of Medical Informatics, 2013. 82(1): p. 10-24.

- Palacios, A.M., L. Sánchez, and I. Couso, *Diagnosis of dyslexia with low quality data with genetic fuzzy systems.* International Journal of Approximate Reasoning, 2010. 51(8): p. 993-1009.

- Kuang, C., et al., *Usher: Improving Data Quality with Dynamic Forms.* Knowledge and Data Engineering, IEEE Transactions on, 2011. 23(8): p. 1138-1153.