**Departamento de Informática – UTFSM**

# Ciclo de Coloquios 2015

El Departamento de Informática de la Universidad Técnica Federico Santa María tiene el agrado de invitar a la comunidad universitaria a un nuevo coloquio a realizarse el día **Martes 04 de Agosto a las 12:00 Hrs** en el Laboratorio de Programación Avanzada (LPA) del Departamento de Informática, Campus San Joaquín. La charla se trasmitirá por videoconferencia al auditorio Claudio Matamoros (F-106) en Casa Central.

## Título

Exploiting Semantic Analysis of Documents for the Domain User

## Invitado

**Prof. Evangelos Milios**
Dalhousie University, Canada
*http://web.cs.dal.ca/~eem/*

## Mini Bio

Evangelos Milios received a Ph.D. in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology in 1986. After spending 5 years doing research in the University of Toronto, he joined York University in 1991 as an Associate Professor. Since July of 1998 he has been with the Faculty of Computer Science, Dalhousie University, where he is Professor and Killam Chair of Computer Science. He has published on the processing, interpretation and use of visual and range signals for landmark-based navigation and map construction in single- and multiagent robotics. His current research activity is centered on modelling and mining of content and link structure of Networked Information Spaces.

## Resumen

Many document organization tasks, such as a student writing the related work chapter of a thesis, a professor surveying the state of the art in a proposal or planning a reading course, or a conference chair organizing sessions would be performed more efficiently through the use of document clustering. Fully unsupervised document clustering does not always yield clusters that are relevant to the user's point of view. In this work, we pursue document clustering algorithms that allow the interactive engagement of the user in the clustering process. The main challenge is how to obtain useful clusters with minimum user effort. To address this challenge, we propose (1) a user-supervised double clustering algorithm, designed in three stages, and (2) a novel approach for mapping documents to entities and concepts. The user-supervised double clustering algorithm was demonstrated to be competitive to state-of-the-art clustering algorithms. It was further extended into an ensemble algorithm to incorporate Wikipedia concepts in the document representation. User supervision was introduced into these algorithms in the form of term supervision (term labelling) and document supervision. A visual interface was designed to make the algorithms accessible to real domain users. The work received the Best Student Paper award at ACM DocEng 2014.

To address the problem of coming up with succinct and intuitive representations of documents in terms of entities and concepts, we have pursued two directions of research: (1) we designed a system that accomplishes entity recognition and disambiguation using the Wikipedia category structure in multiple languages.We are currently extending this system to concept recognition and disambiguation. Our system got the first prize in the ERD challenge at ACM SIGIR 2014; (2) we proposed a simple but very effective approach for computing semantic relatedness between words and documents based on the Google n-gram corpus, which is competitive to human performance on standard word pair data sets. The clustering work is joint with H. Nourashraf and D. Arnold, the ERD work with Marek Lipczak and Arash Koushkestani, and the Google n-gram based semantic relatedness with Aminul Islam and Vlado Keselj.

## Lugar y Fecha

Martes 04 de Agosto de 2015, 12:00 Hrs.
LPA, Departamento de Informática UTFSM San Joaquín
*Sala F-106 (Videoconferencia), DI UTFSM Casa Central*