PODER EXECUTIVO
MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

CONVITE À COMUNIDADE

A Coordenação do Programa de Pós-Graduação em Informática PPGI/UFAM tem o prazer de convidar toda a comunidade para a sessão pública de apresentação de defesa de tese:

Unsupervised Information Extraction by Text Segmentation

RESUMO: In this work we present a new unsupervised approach for the information extraction problem that relies on very effective matching strategies instead of explicit learning strategies to perform the extraction task. Differently from previous proposed approaches that require rely on manual training, ours relies on a knowledge base that can be automatically built using pre-existing data sources, such us Wikipedia, FreeBase, etc. To demonstrate its efficiency, we present here two distinct extraction methods, ONDUX and JUDIE, for the problem of Information Extraction by Text Segmentation (IETS). ONDUX (On Demand Unsupervised Information Extraction) is a new unsupervised probabilistic approach for IETS that relies on the effectiveness of our matching strategy to disambiguate the extraction of certain attributes through a reinforcement step that explores sequencing and positioning of attribute values directly learned on-demand from test data, with no previous human-driven training, a feature unique to ONDUX. The other extraction method, JUDIE (Joint Unsupervised Structure Discovery and Information Extraction) is a new method for automatically extracting semi-structured data records in the form of continuous text (e.g., bibliographic citations, postal addresses, etc.) and having no explicit delimiters between them. While in state-of-the-art Information Extraction methods the structure of the data records is manually supplied the by user as a training step, JUDIE is capable of detecting the structure of each individual record being extracted without any user assistance. This is accomplished by a novel Structure Discovery algorithm that, given a sequence of labels representing attributes assigned to potential values, groups these labels into individual records by looking for frequent patterns of label repetitions among the given sequence. The quality of ONDUX and JUDIE is evaluated through extensive experiments that we report here. These experiments indicate that our proposed approach achieve high quality results and it is able to clearly support information extraction methods in a set of real applications when compared to state-of-the-art information extraction approaches.

CANDIDATO(A): ELI CORTEZ CUSTÓDIO VILARINHO

BANCA EXAMINADORA:

Prof. Altigran Soares da Silva - IComp/PPGI (Presidente)
Prof. Alberto Henrique Frade Laender - DCC/UFMG
Prof. Divesh Srivastava - AT&T/USA
Prof. Edleno Silva de Moura - IComp/PPGI
Prof. Caetano Traina Junior - USP

LOCAL: Sala de seminários do Instituto de Computação

DATA: 17/12/2012

HORÁRIO: 09:00

Professor Dr. Edleno Moura da Silva

Coordenador do Programa de Pós-Graduação em Informática PPGI/UFAM

Av. Rodrigo Otávio, 6.200 • Campus Universitário Senador Arthur Virgílio Filho • CEP 69077-000 • Manaus, AM, Brasil
Tel. (092) 3305 1193 E-mail: secretaria@icomp.ufam.edu.br www.ppgi.ufam.edu.br