MAP-i Doctoral Program in Computer Science

Thesis Proposal

Title:

On Semantically Indexing Digital Documents through an Automatic and Social Classification System

Brief Description:

One of the most used models in semantic classification of documents is the Vector Space Model (VSM). Each document in this model is described as a vector in a *n*-feature space. Although this representation is intuitive and computationally simple, it is expensive from a time-cost perspective, especially for collections of thousands of documents.

Advanced techniques to reduce the time-space complexity of this classification problem have been proposed being the Latent Semantic Indexing (LSI), probably, the most notorious. This technique builds on a *single value decomposition* which conveys a way to implement the LSI efficiently. Recently, have also been proposed variations of clustering methods to profit from the new classification techniques.

In parallel with the developments in automatic classification, systems for social classification have emerged. Currently there are sites which act as repositories of videos, links, texts,… And most of them support some social classification system. Tagging and Star Rating are the most common. However, new forms of providing feedback from an asset/resource are being proposed every day.

The main goal of this work is to study how particular issues like matrix representation, sensitivity and storage issues are affected in a document classifying engine coupled with a repository of clusters of documents. The second goal is to propose new methodologies for integrating an automatic classification of documents with a social classification derived from a folksonomy.

Pre-requisites:

Preference will be given to candidates with solid formation in programming, in mathematics and in the theoretical and practical concepts of design and implementation of Web applications.

Supervisor:

Álvaro Reis Figueira
Departamento de Ciência de Computadores
Faculdade de Ciências da Universidade do Porto
Rua do Campo Alegre, 1021/1055, 4169-007 - Porto
Phone: (+351) 220 402 932 Fax: (+351) 220 402 950
E-mail: arf **** dcc_fc_up_pt
Web: http://www.dcc.fc.up.pt/~arf

Research Unit:

Center for Research in Advanced Computing Systems (CRACS)
Web: http://cracs.fc.up.pt