Progress towards automated ETD cataloging

Venkat Srinivasan[1] and Edward Fox*†[1]

[1]Virginia Tech [Blacksburg] – Virginia Polytechnic Institute and State University Blacksburg, VA 24061-0002, United States

Abstract

Personnel in the Digital Library Research Laboratory (DLRL) at Virginia Tech have been engaged for more than 25 years in developing software to assist comprehension, manageability, and increased adoption of ETDs and their collections. One example was software for automatic generation of concept maps for effective ETD summarization, aimed to assist learning across languages. Taking a cue from this and other such similar efforts at DLRL, and keeping in mind the broader goals of the DLRL to make scholarly knowledge more accessible, we started an initiative in 2008 to develop software to automatically assign topical categories for all the ETDs in the world. The aim was to facilitate browsing and searching of the collection, especially subject-oriented browsing and faceted searching. Further, since many libraries the world over spend substantial amounts of money to catalog (categorize) ETDs, we aimed to assist librarians in this tedious and time-consuming task. Accordingly, we have developed Machine Learning techniques to automatically categorize ETDs into the Library of Congress (LCC) topical taxonomy, which is the dominant categorization scheme used in libraries worldwide. As a prelude to this goal, we developed in 2008 tools to identify science, technology, engineering, and/or mathematics (STEM) ETDs from a given collection of ETDs. Using a testbed of ETDs drawn from four major US universities, we developed software that could identify STEM ETDs with a high degree of accuracy. Subsequently, in an earlier edition of the ETD conference (2008),we reported our results on categorization of ETDs into the (top level nodes of the) DMOZ (Open Directory Project, named from directory.mozilla.org) category system. Using lessons learned from these studies, we started developing improved software for LCC classification of ETDs. This required much deeper analysis, as well as refinement of methods and experimentation to ensure scalability to manage millions of large PDF documents. We first conducted experiments on a small set of ETDs obtained from the NDLTD Union Catalog, in order to demonstrate the feasibility of our methods. In this paper we describe our most recent efforts. We illustrate the substantial progress we have made towards our goal of classifying all available ETDs. We summarize the tools for categorizing ETDs, and highlight the classification results obtained therein. We also present additional insights arising as a consequence - like overall topical trends in ETDs, trends in specific topical areas over time, inter-disciplinarity characteristics with respect to various areas, etc. In the near future, we intend to classify the entire set of ETDs available through the NDLTD's Union Catalog into the LCC. It is hoped that in addition to providing automated tools to libraries to assist the cataloging process, the results would help describe the overall ETD landscape and stimulate further ETD-related research in areas pertaining to knowledge discovery.

On the authors:

*Speaker

†Corresponding author: fox@vt.edu

Venkat Srinivasan is an independent scientist based in Blacksburg, VA. Venkat holds a Bachelor's degree in Computer Engineering from the University of Delhi, India, and a Graduate degree in Computer Science (MS, and ABD en route to PhD) from Virginia Tech, USA. Venkat's primary area of research are Machine Learning and Statistical Inference, and their application to problems in Information Retrieval and the broader area of knowledge and information management. Venkat's graduate research focused on analysis of large collections of Electronic Theses and Dissertations (ETDs) to uncover patterns to aid their automatic classification into the Library of Congress category system. Venkat's research has been published IEEE, IJDL and other journals, and several conferences in the areas of information retrieval and information management. Venkat currently provides scientific consulting services to companies in the US and abroad in the areas of resource optimization, fraud detection, and a range of forecasting problems.