Progress towards automated ETD cataloging

Venkat Srinivasan [*1] and Edward A. Fox [†2]

[1,2]Department of Computer Science, Virginia Tech, Blacksburg, VA, USA

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 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 to 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 (ML) 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

[*]svenkat@vt.edu
[†]fox@vt.edu

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 through the NDLTD Union Catalog, in order to demonstrate the feasibility of our methods.

In this paper we describe our most recent efforts. 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 English ETDs readily 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.

1 Introduction

Libraries around the world spend substantial amounts of money in cataloging ETDs. Since most libraries use either the Library of Congress Classification (LCC) or the Dewey Decimal Classification (DDC) system, the librarians and catalogers are expected to be closely familiar with these systems, as well as have substantial subject matter knowledge in topics the ETDs deal with in order to accomplish the categorization task. External knowledge bases like WorldCat's book look-up service [1] that provide the appropriate LCC or DDC categories for books and other manuscripts are not as readily usable for ETDs, since ETDs typically are local (e.g., an ETD submitted to a particular university, that finds its way to the university library), and are thus unlikely to have been cataloged elsewhere.

To assist catalogers in this task, we have developed software that for a given ETD would automatically assign the most suitable LCC categories. Our techniques are based on Machine Learning (ML) principles which have been used for over two decades to categorize text, but so far have had only limited applicability towards categorizing book-length documents.

As a first step, we had developed software in 2009 to automatically distinguish STEM ETDs from non-STEM ones. We achieved a fairly satisfactory accuracy rate, and published our results [5]. As a follow-up to this work we developed tools to classify ETDs into broader categories based on the DMOZ category system [2] and presented the results at an earlier ETD conference [6].

Presently, our software can assign ETDs into one of 70 chosen leaf level categories drawn from the LCC (please see the Appendix at the end of this

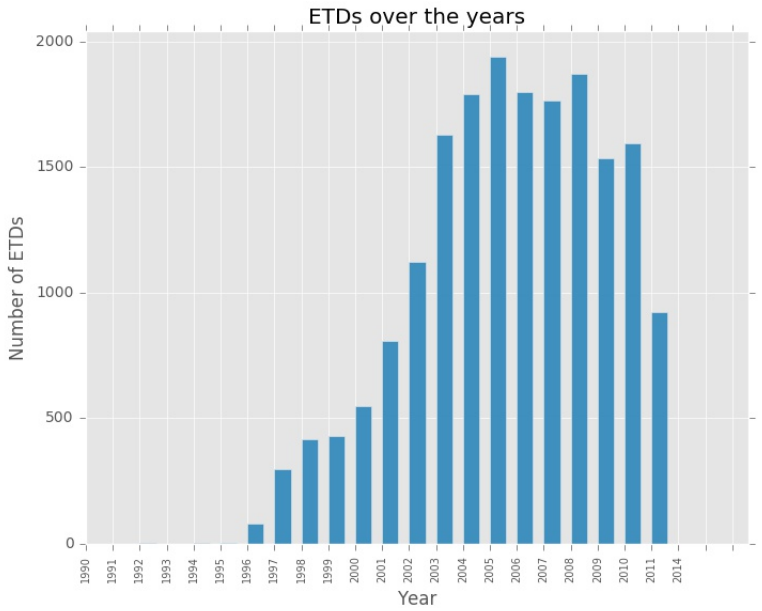

Figure 1: ETDs over time

paper for the list of categories). LCC of course is a full-fledged category tree that has several thousand nodes and a few hundred leaf level nodes. As we proceeded with the categorization task though, we observed that most ETDs tend to fall into a select few "major" topical categories, and "child" (or leaf) categories thereof that were just about 1-2 levels below the "major" category in the tree. Typical ML algorithms first classify documents into major categories and then into minor categories and so on. Since in our case, the catalogers would benefit from assigning an ETD into the most specific category possible as opposed to a general higher level category, we chose only the appropriate leaf categories for our classification experiments, and developed techniques that would directly map the ETDs to these categories. The Appendix lists the specific categories that were selected and the number of ETDs mapped to them for training purposes (please see Section 2 for details).

In the next section, we summarize our methods and results. We then conclude by making some general observations, and laying out guidelines for possible future research.

2 ETD Categorization

Our methods are based on principles derived from hierarchical text classification methods from the domain of supervised machine learning. We "train" our

algorithms to recognize statistically significant keywords occurring in different parts of ETDs - title, abstract, beginning chapters, etc., and to use those as cues to predict specific LCC categories. We have a collection of $\sim$19000 ETDs from 4 major US universities that have been mapped to the 70 leaf level LCC categories by a Virginia Tech librarian. We used this collection to *train* our methods based on a specific ML technique known as Support Vector Machines (SVMs) to zero in on important keywords and phrases for each category. The presence/absence of these keywords in the ETDs (metadata and/or full text) would indicate the category of the ETD. The ETDs that we used for training purposes had been submitted to the host institutions starting 1997. Their distribution over time is shown in Figure 1.

The input to our software is the ETD metadata, particularly the title, abstract and keywords. It is highly preferred to have the full text of the ETDs also be fed as input, but this is not a strict necessity as our methods can perform well with just the metadata. One of the distinguishing features of our algorithms when compared with the existing ones is their ability to make judicious use of keywords and phrases occurring in the body of the ETD and combine this information with keywords in the metadata in an efficient fashion to eke out greater categorization accuracy. Once the ML *training* in such a fashion has been done, the software is ready to ingest incoming ETD metadata (and full text if available) for classification, and then outputs, for each ETD, the most suitable LCC leaf category.

Our methods have an accuracy of over 70% at identifying the category of an ETD. This is considerably higher than the existing state-of-the-art ML algorithms. The running time of our algorithm is $\sim$2hrs. This is in fact the time it takes to complete the ML *training*. The classification of a new, unseen ETD into its LCC category is near instantaneous. Our software has been programmed in the Python programming language and is available for general use on request.

We have also developed several other tools as by-products of our work, that could be of interest to the broader LIS community. Primarily, these tools help with parsing of the PDF files (ETDs are typically submitted as PDFs), extracting targeted information like table of contents, individual chapters, tables/figures and/or their captions, and bibliography, etc. We have developed tools that can process and extract to varying degrees of accuracy several of these elements from the body of the ETDs. A pilot study and some early results were published in an earlier ETD conference [7]. Some of our more recent results in were published in [4].

In addition, we have developed several other tools that allow for monitoring of specific topical areas and sub-areas that a cataloger may be interested in tracking. These tools complement the categorization tools quite well. For

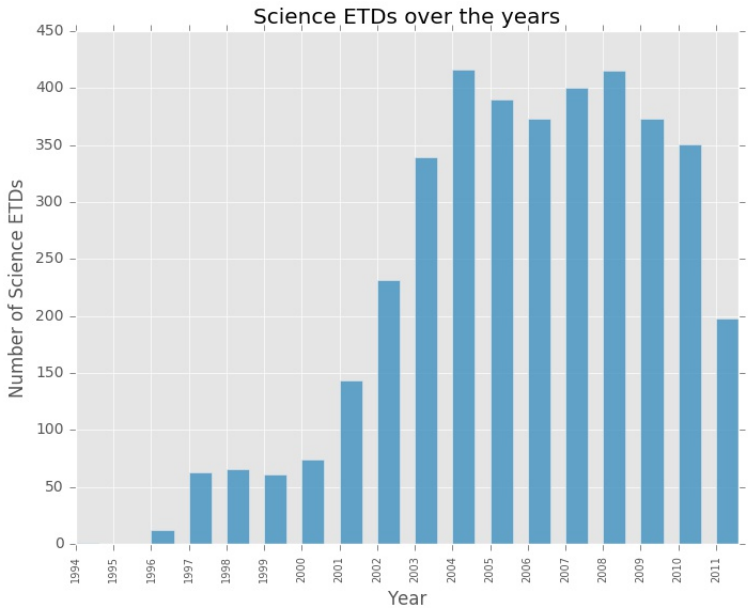

Figure 2: Science ETDs over time

example, a cataloger could use them to study the evolution of a particular topic in ETDs over time for analytics and reporting purposes, besides just cataloging ETDs. One sample use case is shown in Figure 2 which shows the number of ETDs that have been submitted in the "SCIENCE" leaf node of the LCC tree over time that were present in our collection. Another sample use is demonstrated in Table 1 which shows the 10 major themes/topical areas in the category "EDUCATION" of the LCC subtree over the years. The keyphrases listed were identified by our SVM based algorithm as the most defining ones for this particular category.

3 Discussion

In this paper we described our tool for performing automated classification of ETDs. The tool could aid professional catalogers in performing their tasks more efficiently and accurately. Our next step is to run the tool using the entire Union Catalog, to assign LCC labels to ETDs present there, results of which could be directly used by libraries.

We also described several other auxiliary tools for performing different kinds of analytics on collections of ETDs. These are useful for topic tracking, targeted information extraction from ETDs, etc. Extraction of individual segments from ETDs, like individual chapters, figures etc. could potentially be useful in de-

Year	Key themes
2004	community schools, moral judgment, school culture, leadership behaviors, learner centered, community college, student government, community colleges, teaching style, teaching styles
2005	high school, vocational education, public administration, school graduates, non vocational, degree programs, associate degree, adult education, colleges universities, junior colleges
2006	special education, global mindedness, administrative support, study abroad, cultural competence, education teachers, higher education, significant difference, cultural identity, abroad group
2007	pre release, programs days, group students, performance periods, students attended, transition process, cclc programs, release handbook, law abiding, intervention group
2008	domain knowledge, student achievement, grievance arbitration, teaching oriented, oriented institutions, job satisfaction, research oriented, educational leaders, capital appropriations, working conditions
2009	file sharing, problem solving, athletic training, training education, sexuality education, solving confidence, fourth year, research question, displaced workers, white fraternities
2010	spiritual quest, stem fields, community service, student athletes, diversity related, novice otas, success persistence, related experiences, state university, civic education

Table 1: Key themes in EDUCATION related ETDs

signing personalized learning modules for users, wherein information tailored for specific users could be drawn from different ETDs and presented to users according to their interests and/or learning needs This is similar to the work by Liang et al.[3] for books, but in the context of ETDs. Several initiatives in DLRL have also addressed the issue of personalized learning for users. Our tools could be a valuable counterpart to such initiatves as well, applied to the domain of ETDs.

References

[1] OCLC WorldCat: Window to the world's libraries. `http://www.oclc.org/worldcat/`. Retrieved Aug 2012.

[2] Open directory project. `http://www.dmoz.org`, Retrieved Mar 2010.

[3] Chen Liang, Shuting Wang, Zhaohui Wu, Kyle Williams, Bart Pursel, Benjamin Brautigam, Sherwyn Saul, Hannah Williams, Kyle Bowen, and C. Lee Giles. Bbookx: An automatic book creation framework. In *Proceedings of the 2015 ACM Symposium on Document Engineering*, DocEng '15, pages 121–124, New York, NY, USA, 2015. ACM.

[4] Venkat Srinivasan and Pranav Angara. *'Classification'. A chapter appearing in 'Digital Library Technologies: Complex Objects, Annotation, Ontologies, Classification, Extraction, and Security'*. Morgan & Claypool, 2014.

[5] Venkat Srinivasan and Edward A. Fox. Global S&T assessment by analysis of large ETD collections. *IEEE Intelligent Systems (special Issue on 'AI and Global Science and Technology Assessment')*, 24(4):68–88, July 2009.

[6] Venkat Srinivasan and Edward A. Fox. Topical categorization of large collections of electronic theses and dissertations. *12th International Symposium on Electronic Theses and Dissertations*, June 2009.

[7] Venkat Srinivasan, Mohamed Magdy, and Edward A. Fox. Enhanced browsing system for electronic theses and dissertations. *14th International Symposium on Electronic Theses and Dissertations*, Sept 2011.

Appendix

Node	Number of ETDs
Agriculture (General)	101
Anthropology	93
Architecture	577
Biotechnology	85
Botany	57
Cattle	50
Chemical engineering	350
Communities. Classes. Races	85
Economic history and conditions	61
Economic theory. Demography	181
Educational psychology	201
Electronics	1477
English language	590
Environmental Sciences	80
Environmental engineering	143
Food processing and manufacture	308
Forestry	321
General Mathematics	448
General Physics	324
General EDUCATION	272
General Including alchemy	535
General Microbiology	60
Geography (General). Atlases. Maps	62
Geology	129
Higher education	51
Horticulture. Horticultural crops	157
Human ecology. Anthropogeography	114
Immunology	36
Industries. Land use. Labor	170
Instruments and machines	977
Landscape gardening. Landscape architecture	88

MUSIC	660
Manufactures	100
Materials of engineering and construction. Mechanics of materials	298
Mechanical engineering and machinery	1271
Mechanics of engineering. Applied mechanics	244
Medicine (General)	114
Meteorology. Climatology Including the earth's atmosphere	126
Mining engineering. Metallurgy	123
Modern languages. Celtic languages	60
Motor vehicles. Aeronautics. Astronautics	110
Natural history - Biology	592
Nuclear engineering. Atomic power	98
Nursing	57
Oceanography	60
Organic chemistry	82
PHILOSOPHY. PSYCHOLOGY. RELIGION	1260
Pests and diseases	73
Philology. Linguistics	226
Physiology	120
Political institutions and public administration	184
Political science (General)	80
Poultry. Eggs	146
Public health. Hygiene. Preventive medicine	54
Recreation. Leisure	126
School administration and organization	512
Science (General)	106

Sociology (General)	202
Special aspects of education	147
Statistics	229
Systems engineering	165
Teaching (Principles and practice)	133
Technology (General)	185
Telecommunication Including telegraphy, telephone, radio, radar, television	39
The family. Marriage. Women	103
Theory and practice of education	1175
Toxicology. Poisons	56
Veterinary medicine	129
Visual arts	240
WORLD HISTORY AND HISTORY OF EUROPE,	386
Zoology	315
TOTAL	**18569**