ETD 2016 Conference, France, July 2016

Progress towards Automated ETD Cataloging

Venkat Srinivasan, Edward A. Fox

Virginia Tech, Blacksburg, VA

{svenkat, fox}@vt.edu

Outline

- Motivation
- Background
- Earlier Results
- ETD Categorization into LCC
- ETD Tools
- ETD Tools Use Cases
- Future Work

Motivation

- ETD cataloging is time intensive and laborious
  - Familiarity with Library of Congress Classification (LCC) and/or Dewey Decimal Classification (DDC) required -- with hundreds of nodes
  - Familiarity with the subject matter of ETDs required -- often highly specialized
  - Supplementary resources like WorldCat look-up do not provide topical categories for ETDs

Motivation

- Automated cataloging tools can result in substantial time and cost savings
  - Automated classification of text into categories -- for over two decades
  - Machine Learning (ML) based tools available
  - But limited applicability and generalizability when applied to long documents like books, ETDs
- Users can benefit from faceted browsing and searching of collections of ETDs

Background

- Digital Libraries Research Lab (DLRL) @ Virginia Tech involved in ETD related research for over 25 yrs
  - Institutional repositories, Open Archiving, domain specific digital libraries (PhysNet), concept maps, …

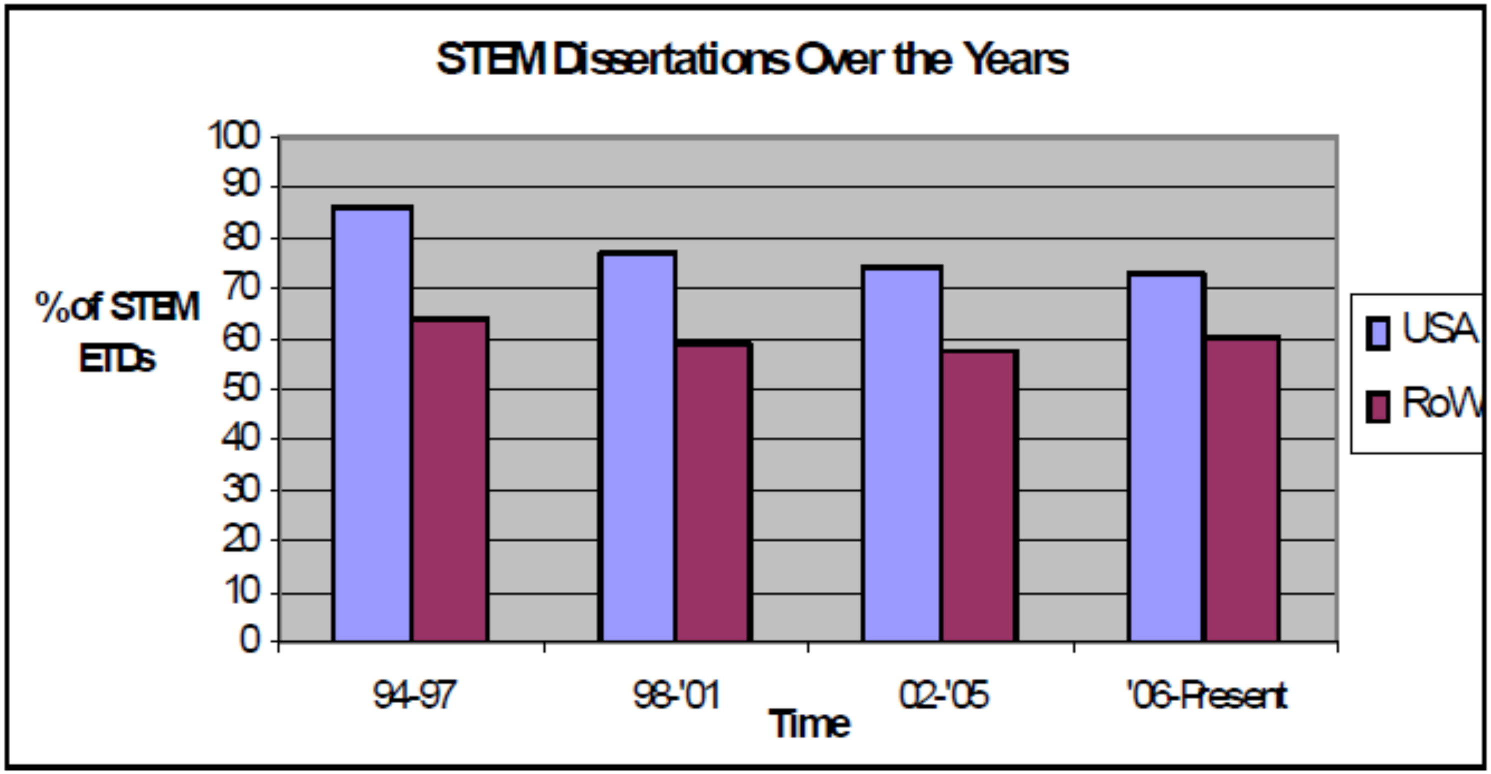
- ETD cataloging project @ DLRL started in 2008
  - Initially: extraction of keywords and phrases to generate concept maps
  - Later: moved towards ETD repository building and cataloging

Earlier Results

- STEM vs. Non-STEM ETD identification
  - (IEEE IS 2009)
  - ML based tools developed to categorize ETDs based on Dublin Core metadata
  - Accuracy of over 90% achieved on a collection of over 100K ETDs

Earlier Results

- STEM vs. Non-STEM ETD identification
  - STEM productivity over time (figure from Srinivasan, Fox IEEE IS 2009)



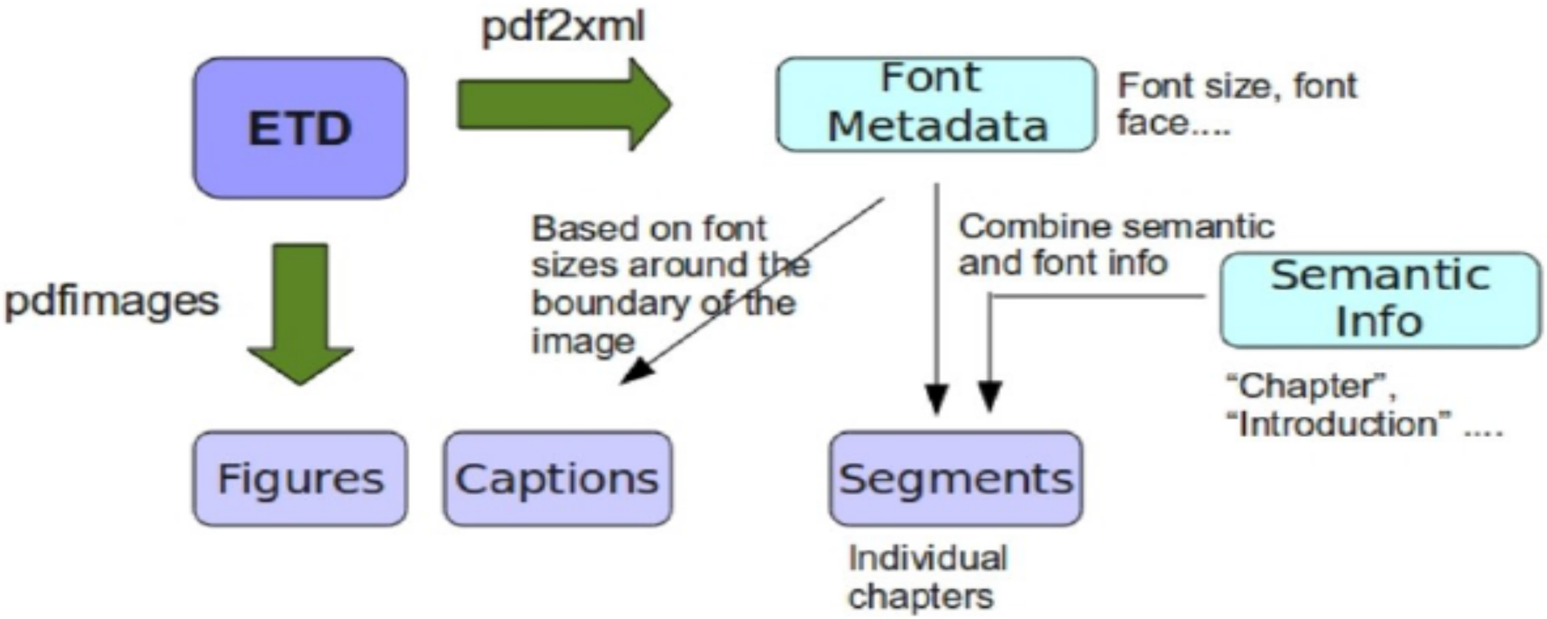
Earlier Results

- ETD classification into DMOZ top level nodes
  - (ETD 2009 Conf.)
  - Broad categories like Business, Health, Society
  - Further tools developed for crawling/harvesting and categorization of ETDs
  - Valuable lessons learned regarding categorization into topical taxonomies
  - ~55K ETDs from 8 major US universities categorized:

Topic	Arts	Business	Computers	Health	Science	Society
K ETDs	4	4	17	4	24	2

Earlier Results

- Faceted Browsing of ETDs (ETD 2011 Conf.)
  - Tools for extracting targeted information from within ETDs (images, chapters, TOCs, etc.)
  - Browsing interface based on all this.


pdf2xml
ETD
Font Metadata
Font size, font face….
pdfimages
Based on font sizes around the boundary of the image
Combine semantic and font info
Semantic Info
"Chapter", "Introduction" ….
Figures
Captions
Segments
Individual chapters

ETD Categorization into LCC

- Virginia Tech Librarian mapped 18K ETDs into LCC
    - “Department Name” field of metadata was mapped into an LCC node
    - ETDs fell into a limited subset of LCC areas

- LCC taxonomy pruned
    - 70 “leaf” nodes selected (mostly levels 2 and 3)

ETD Categorization into LCC

- Machine Learning (ML) techniques for cataloging
    - “Learn” the characteristics of each node
        - e.g., representative keywords that occur often in one node but not in another
    - Successfully used in other text categorization, but not so much for book-length documents
    - Our approach combines metadata information with ETD segments (table of contents, front matter, etc.) to develop faster and more accurate techniques

ETD Categorization into LCC

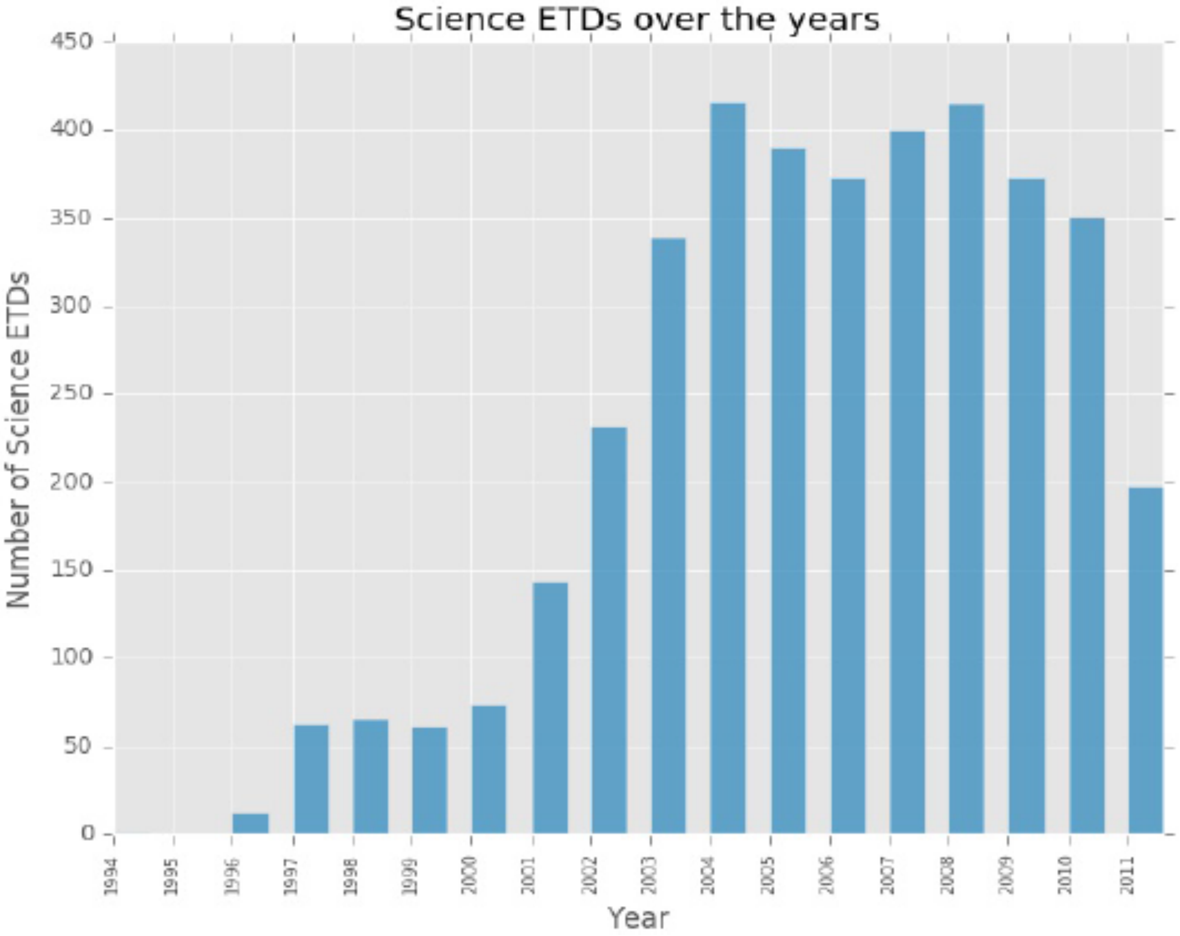
- Results
  - Learning and testing done in <2hrs
  - Operation is easy:
    - Point to folder containing ETDs in PDF
    - Outputs LCC category node for each ETD
  - ~70% accuracy
    - classifying 18K ETDs into 70 nodes

ETD Tools

- Other tools
  - Extracting individual chapters, table of contents info, figures and captions, etc.
  - Varying degrees of accuracy
    - chapter identification ~75%
    - figure extraction ~40%
  - Commercial tools (vs. our open source tools) likely to result in greater accuracy

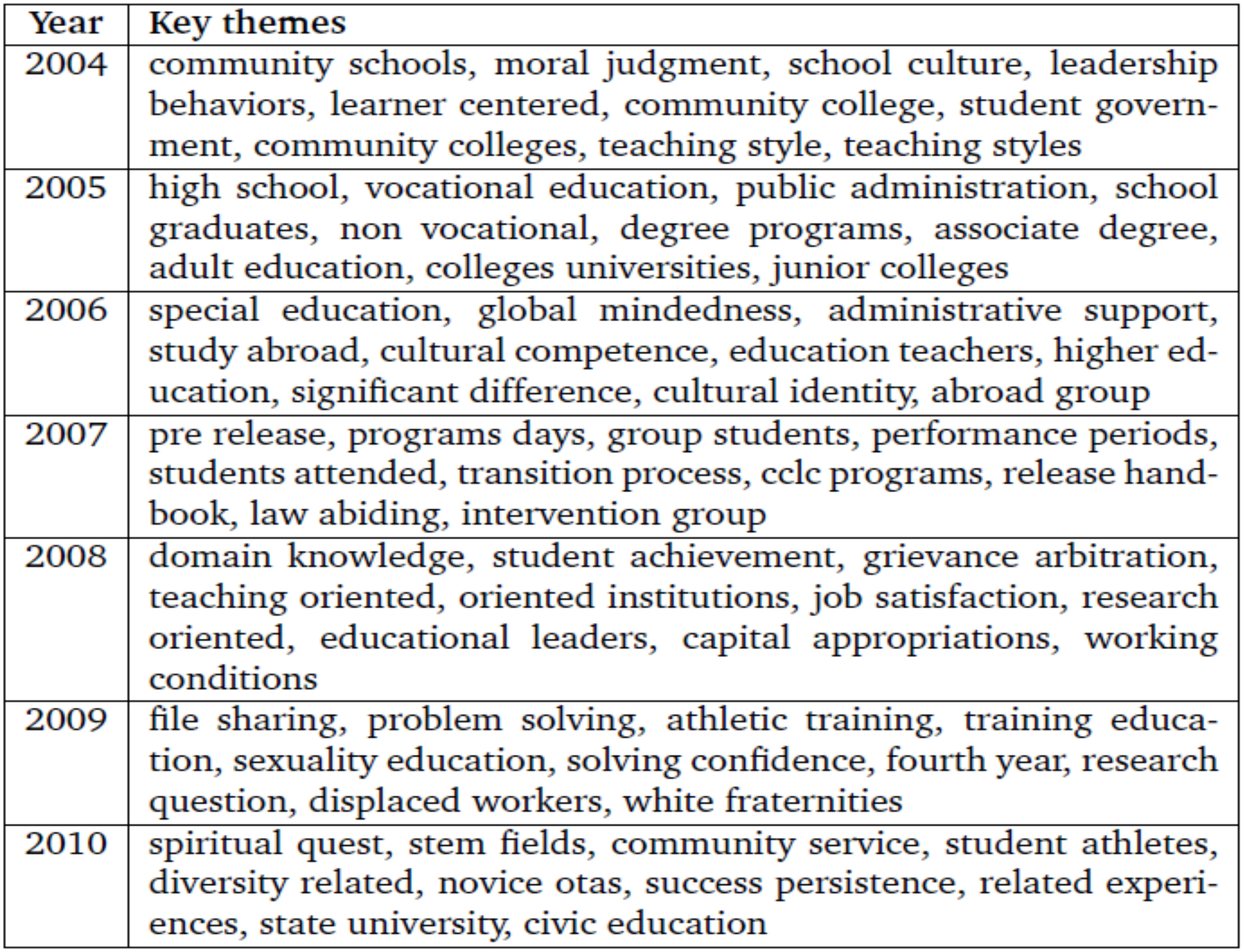
ETD Tools Use Cases

- Science ETDs in our collection, over time



ETD Tools Use Cases

- Major themes in the “Education” leaf of LCC over time



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

Future Work

- Categorize the entire Union Catalog ETDs into LCC
- Explore applications to personalized digital libraries
    - E.g., leverage BbookX
- Use commercial tools for segmentation and extraction
- User Interface design for browsing, searching ,etc.

Future Work

- Applications and studies
    - Topical trend analysis
    - General purpose ETD analytics
    - Complementing and extending existing ETD tools (e.g., Concept Maps)

References

1. Ryan Richardson and Edward A. Fox: Using Concept Maps in NDLTD as a Cross-Language Summarization Tool for Computing-Related ETDs. In Proc. ETD 2007, Uppsala, Sweden, June 13-16, 2007, http://fox.cs.vt.edu/talks/2007/20070613ETDfoxRichardsonCmaps.ppt
2. Ryan Richardson, "Using Concept Maps as a Cross-Language Resource Discovery Tool", July 2007, PhD dissertation, http://scholar.lib.vt.edu/theses/available/etd-07022007-184525/
3. Venkat Srinivasan, Edward A. Fox. Global Science and Technology Assessment by Analysis of Large Collections of Electronic Dissertations, in Hsinchun Chen, Ronald N. Kostoff, Chaomei Chen, Jian Zhang, Michael S. Vogeley, Katy Borner, Nianli Ma, Russell J. Duhon, Angela Zoss, Venkat Srinivasan, Edward A. Fox, Christopher C. Yang, Chih-Ping Wei, "AI and Global Science and Technology Assessment," IEEE Intelligent Systems, 24(4): 68-88, July/August, 2009 http://www.computer.org/portal/web/csdl/magazines/intelligent#4
4. Venkat Srinivasan and Edward Fox. Topical Categorization of Large Collections of Electronic Theses and Dissertations, Proc. ETD 2009, Pittsburgh, PA, June 10-11, 2009, http://fox.cs.vt.edu/talks/2009/20090611ETD2009ClassifyTalk.ppt
5. Srinivasan, Venkat, Magdy, Mohamed, and Fox, Edward. Enhanced Browsing System for Electronic Theses and Dissertations. In Proc. ETD 2011 - 14th International Symposium on Electronic Theses and Dissertations. Cape Town, South Africa. September 13-17, 2011

Comments?

Questions?