Topical Categorization of Large Collections of Electronic Theses and Dissertations

Venkat Srinivasan & Edward A. Fox
Virginia Tech, Blacksburg, VA, USA
ETD 2009 – June 11, 2009
Outline

- Introduction
- Goals
- Approach
- Results
- Future Work
Introduction – great source

- Electronic submission of dissertations is increasingly preferred.

- ETDs are a great information source.
  - Substantial amount of research on a topic
  - Thorough literature review
  - Pointers to other resources (Reference section)
Introduction - under-utilized

- Yet ETDs are under-utilized.
  - Research papers, books etc. are still major (and in some cases the only) sources of information for most people.
  - Most people (except grad students trained in this) don’t even think about reading a dissertation!

- Possible causes
  - Access to ETDs not streamlined.
    - Users don’t know where to look for ETDs.
    - ETDs of interest could be buried in search engine results.
  - Some universities do not allow outside access to their ETD collection.
Introduction - needs

- Efforts have been made to make ETDs more accessible.
  - NDLTD, VTLS, Scirus, etc. provide means of access to ETDs from different universities.

- Not very feature rich and convenient:
  - Users search for ETDs based on keywords.
  - Don’t know what lies underneath (no idea about the size, topical coverage, etc. of ETD collections)
  - Not very amenable to browsing (users have to sift through search results)
Goals

- Provide a portal to ETD collections of more different universities
- Provide value added services
  - Categorize by topic
  - Support searching and browsing the collection using various criteria (by topic, keywords, date, author, etc.)
Goals - priorities

- Set up infrastructure for crawling ETDs of various universities
- Come up with techniques for categorizing them into topical areas
- Set up a user-friendly search and browse interface
Approach

- Crawl ETDs from various universities
- Develop a taxonomy
- Categorize ETDs into topics in the taxonomy tree
- Index the ETDs
- Develop a search and browse interface
Approach - crawling

- NDLTD’s Union Catalog - as starting point
- Dublin Core metadata gathered
- URLs used to crawl ETDs and other data from the respective universities’ websites
- Custom crawlers written
  - Technologies used: Perl, and other open source Perl libraries (WWW, Mechanize, etc.)
  - All metadata (Dublin Core metadata from Union Catalog, and the metadata obtained from respective universities) is stored in our MySQL backend database.
Approach - taxonomy

- Need medium generality and specificity, as opposed to those from Proquest, DMOZ, or Wikipedia.
- For example, DMOZ has more than 500,000 nodes!
- Solution?
  - Prune the DMOZ category tree, and then enhance it using Proquest categorization system.
Approach – taxonomy levels

Category Tree (only top 2 levels shown)
Approach - categorize

- Supervised classification approach used
- Training set built by using topic labels as query to Google
- 50 webpages retrieved and used for training Naïve Baye’s classifier for each node (to distinguish between its children)
- ETD metadata used for categorization
- Level-wise categorization
Category Tree

- Category label for each node used as query

Google

- Top 50 webpages (for each node in the tree)

Document Sets

- Cleanup (stemming, stopword removal, etc.)

ETD Collection

- ETD metadata used for categorization

Naïve Bayes Classifiers

- Level-wise categorization

Training Sets

- Training

Categorized ETDs

- Browsing

ETDs categorized into a node of the category tree (after classification)

Algorithm Pipeline
Results

- Crawled metadata for all the ETDs from the NDLTD Union Catalog
  - ~800,000 ETDs in Union Catalog
  - 15 Dublin Core fields extracted and stored
- Crawled ~200,000 dissertations from the respective universities (where permissible) and indexing is in progress
  - Technology used: Lucene search engine
- More dissertations being crawled
Results (contd.)

- Enhanced taxonomy developed
- Some subtrees are shown in the following few slides.
- The taxonomy currently is 4 levels deep and has ~200 nodes.
- It is being enhanced to be 5-6 levels deep.
Enhanced taxonomy (some nodes from the “Arts” subtree shown)
Results (contd.)

- Enhanced taxonomy (some nodes from the “Business” subtree shown)
Results (contd.)

- Categorized >74K ETDs from 8 universities
  - MIT, Virginia Tech, Caltech, NCSU, Georgia Tech, Ohiolink, Rice, Texas A&M
- Categorized into 5 topical areas (Arts, Business, Computers, Health, Science, Society)
- Categorization into lower levels of category tree (levels 3 and 4, that is) is in progress
## Results (contd.)

<table>
<thead>
<tr>
<th>Name of the University</th>
<th>Total No. of ETDs</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Arts</td>
</tr>
<tr>
<td>MIT</td>
<td>29804</td>
<td>653</td>
</tr>
<tr>
<td>Virginia Tech</td>
<td>11976</td>
<td>742</td>
</tr>
<tr>
<td>Ohiolink</td>
<td>8020</td>
<td>1056</td>
</tr>
<tr>
<td>Rice</td>
<td>6685</td>
<td>937</td>
</tr>
<tr>
<td>NCSU</td>
<td>5026</td>
<td>283</td>
</tr>
<tr>
<td>Texas A&amp;M</td>
<td>4834</td>
<td>302</td>
</tr>
<tr>
<td>CalTech</td>
<td>4774</td>
<td>58</td>
</tr>
<tr>
<td>Georgia Tech</td>
<td>3582</td>
<td>32</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>74701</strong></td>
<td><strong>4063</strong></td>
</tr>
</tbody>
</table>
Results (contd.)

- Algorithm is time efficient.
  - Training the classifier is done offline.
  - Classification is fast.
  - Classifying this collection of ~74,000 ETDs took <30 mins.
- Hopefully classifiers developed can be applied to other data and in other systems.
Future Work

- Increase coverage
  - Crawl more ETDs
  - Collaborate with universities and consortia to gain access to ETD collections
- Better categorization approaches
  - Leverage query expansion techniques to build training set
- Web interface to facilitate browsing and search
- User studies to measure the efficacy of the system
Questions ?

svenkat@vt.edu
fox@vt.edu

Demo info available at
http://fox.cs.vt.edu/etdbrowse/