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

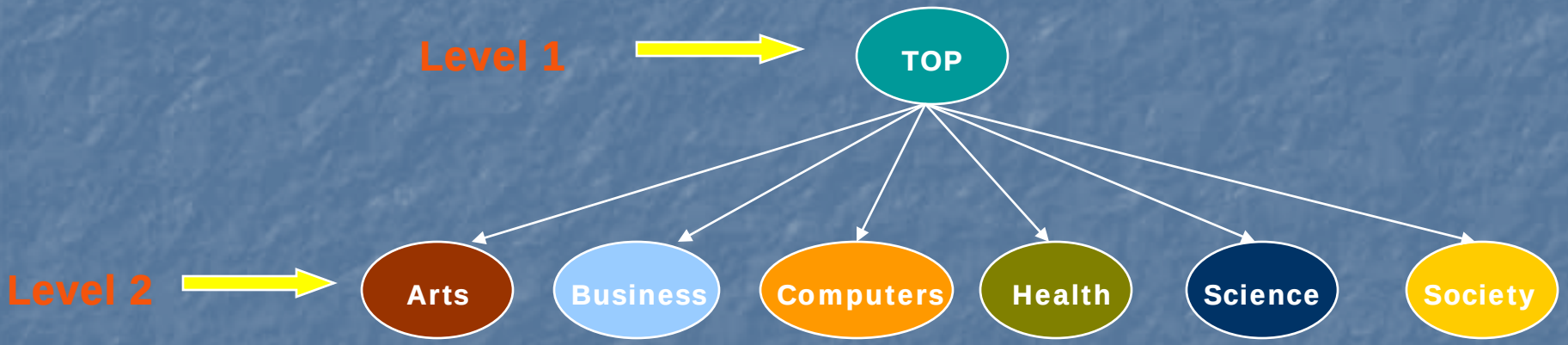
- NDLTD's Union Catalog - as starting point
- Dublin Core metadata gathered
- URLs used to crawl ETDs and other data from the respective universities' websites
- Custom crawlers written
  - Technologies used: Perl, and other open source Perl libraries (WWW, Mechanize, etc.)
  - All metadata (Dublin Core 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



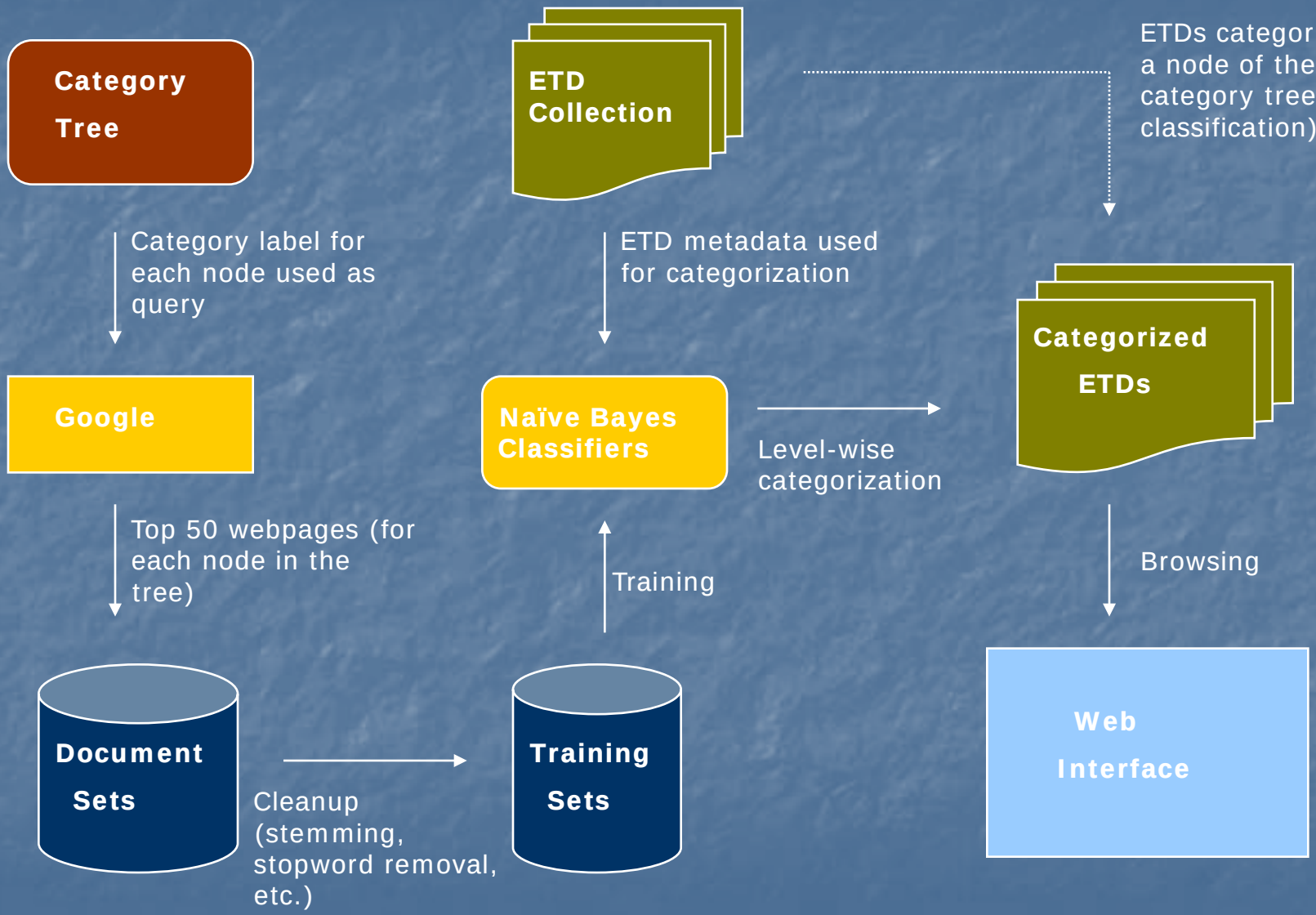
Category Tree (only top 2 levels shown)

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

Approach (contd.)

Algorithm Pipeline



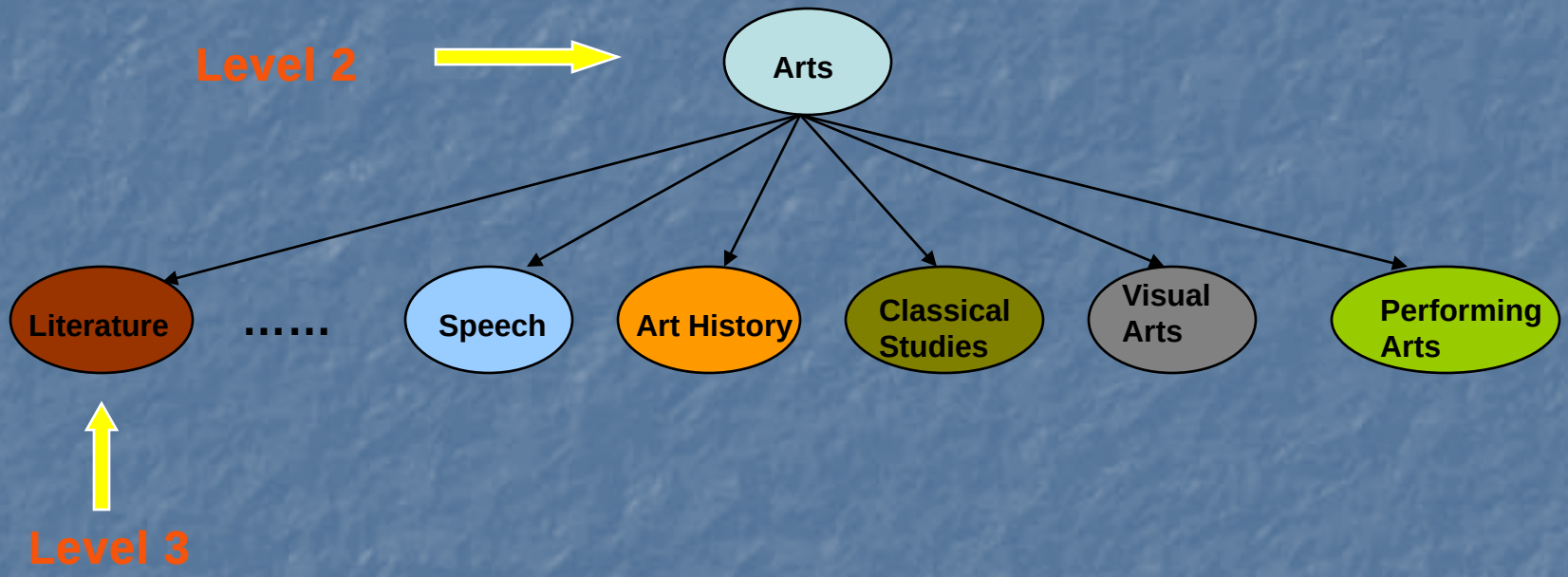
Results

- 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.)

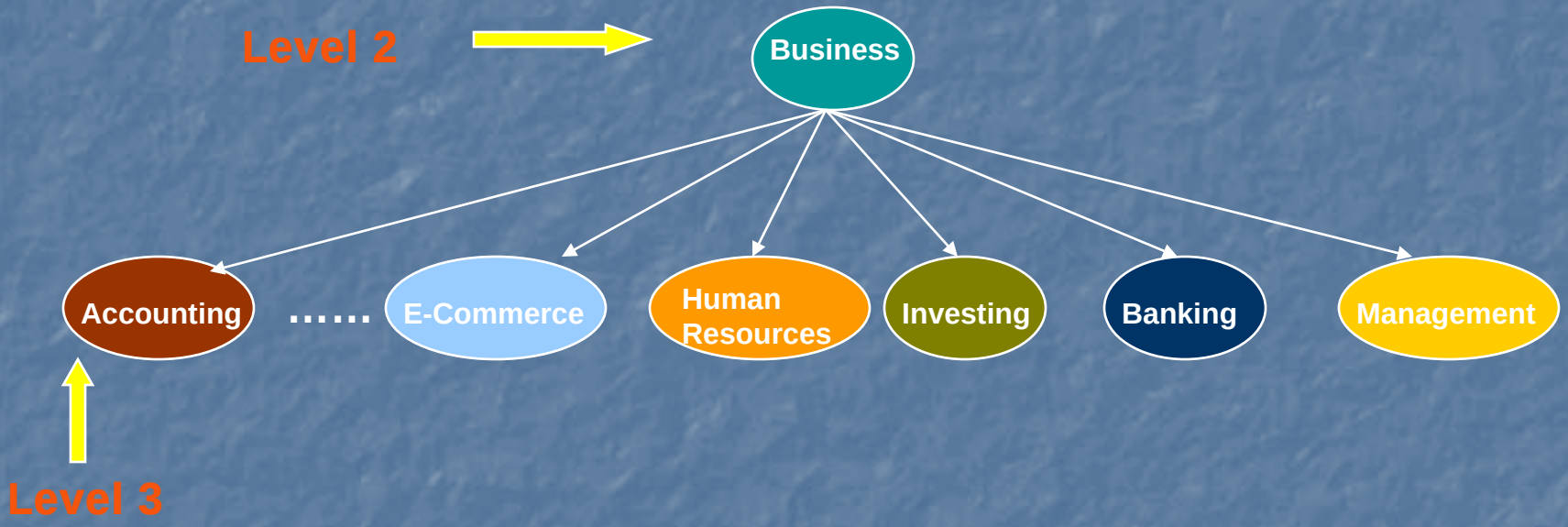
- Enhanced taxonomy developed
  - Some subtrees are shown in the following few slides.
  - The taxonomy currently is 4 levels deep and has ~200 nodes.
  - It is being enhanced to be 5-6 levels deep.

Results (contd.)



- 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.)

Name of the University	Total No. of ETDs	Category					
		Arts	Business	Computers	Health	Science	Society
MIT	29804	653	1847	6507	375	7141	555
Virginia Tech	11976	742	627	2665	1218	3317	340
Ohiolink	8020	1056	350	1267	1322	2887	345
Rice	6685	937	235	1181	145	2412	62
NCSU	5026	283	245	1419	512	2436	114
Texas A&M	4834	302	363	1363	566	2115	125
CalTech	4774	58	52	1392	29	3096	18
Georgia Tech	3582	32	133	1348	85	1233	23
TOTAL	74701	4063	3852	17142	4252	24637	1582

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/