ETD 2009

Topical Categorization of Large Collections of Electronic Theses and Dissertations

Venkat Srinivasan, Edward A. Fox
Department of Computer Science
Virginia Polytechnic Institute and State University
{svenkat, fox}@vt.edu

Electronic Theses and Dissertations (ETDs) form an important part of scholarly work. Many universities in the USA, and other parts of the world, require their students to submit their theses and dissertations in electronic form. The ETDs are hosted by the respective universities, and no single point of access exists to the different ETD collections. Various initiatives like NDLTD have aimed to provide a unified mode of access to the ETD collections of different universities. Currently, however, users can only search (using various web interfaces) for ETDs, with minimal support for browsing. Many would like to browse by topic, using some appropriate category scheme, and to obtain information such as size, for parts of a hierarchical classification related to topical coverage.

In this paper we address some of these issues. We provide a portal to ETDs from different universities, by harvesting/crawling the ETDs wherever permissible. We also have developed a topical taxonomy derived from DMOZ, and approaches to categorise ETDs into that taxonomy. We present categorization results for ETD collections of some universities in the USA that have large numbers of ETDs in the Union Catalog.

1. Introduction

Preparation and submission of dissertations in electronic form has become an increasingly preferred way in many universities around the world. Initiatives like NDLTD[1] have made a concerted effort to provide wider access to, and dissemination of, ETDs. NDLTD's Union Catalog, for example, currently contains metadata and URLs for more than 700,000 ETDs in various languages and from different parts of the world.

Such large collections of ETDs however also need to be complemented by efficient and user friendly modes of access. In the NDLTD website for example, the only interfaces that exist are search and browse based. Two such interfaces are those provided by VTLS[2]and Scirus[3] The Scirus interface allows the users to search based on the metadata, like title, author name, year of publication, etc. The Scirus interface also allows the users to search within broad topical areas such as Mathematics, Physics, etc. The VTLS interface allows users to search dissertations by language, in addition to some metadata fields like year of publication, title, etc., and then the users can refine the search results based on additional keywords.

Such methods of access however give the users no idea about the size per topic, and the topical coverage of the available content, and clearly presuppose that the users have a good idea about the kind of information they are looking for. Much of the information seeking under such circumstances involves the user sifting through results that are not of interest or just related tangentially. Clearly this places a serious burden on information seekers, especially those not knowing the terminology for an area of interest.

A valuable service to add to such large collections of ETDs would be to categorize them into different topical areas, and allow the users to browse by topic. In this paper we present our work relating to categorization of a collection of ETDs harvested from NDLTD's Union Catalog. We further enhance this service by providing additional options like searching within categories, browsing by keywords, dates, etc.

2. Browsing by Topics

One of the approaches to implement the "browse by topic" scheme for ETDs is to make use of keywords associated with the ETDs, and allow the users to browse by keywords. The user could then browse the collection, say by navigating the tag cloud generated using the keywords. This approach is unsuitable for browsing an ETD collection for many reasons. Firstly, not all submitters of ETDs necessarily supply keywords for their ETD. This in fact happens to be the case with the ETDs in the Union Catalog. A significant number of them do not have keywords associated with them. Secondly, different users often use different terms to describe the same concept. This leaves open the possibility of ETDs getting marked with different keywords despite having identical topical coverage.

Hence, in order to categorize ETDs into topical areas, we have developed a taxonomy based on the **Mo**zilla **D**irectory or DMOZ[4] category tree. A taxonomy is a hierarchical organization of concepts, and relationships between them, in a particular domain. Taxonomies have been used in many areas like biology, physics, etc. to assign domain specific entities (in our case, ETDs) into corresponding categories. This not only facilitates browsing, but entities that are similar, also implicitly get grouped into the same category. Figure 1 shows the (pruned) top 2 levels of the

DMOZ category tree. In this paper we report categorization results for the second level of the (pruned) DMOZ category tree.

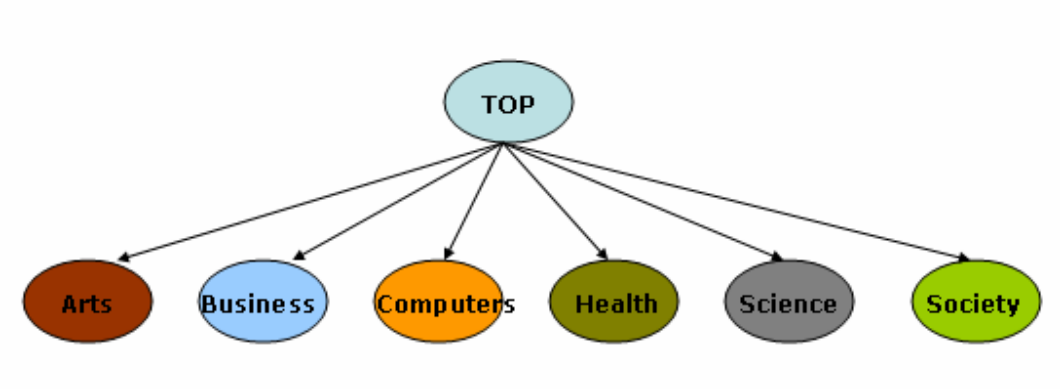

Figure 1. DMOZ top level nodes after some pruning

3. Methods

ETD categorization under these settings is a multi-step process. Our approach to ETD categorization is described in Figure 2. Various intermediate steps involved are described in detail below.

3.1 Building a Taxonomy

In the first step we build a suitable taxonomy for ETDs. While there exists at least one existing taxonomy for ETDs (viz. the one developed by Proquest[5]), we found it to be unsuitable for our purposes for various reasons. Firstly, the taxonomy is very general and not deep enough to cover specific categories within a domain. A taxonomy that is only 2 levels deep and has say "Computer Engineering" as the most specific category is unlikely to be of much help in browsing. Similarly a taxonomy that is say 10 levels deep and is very specific, is also unlikely to be of much help.

We have made use of an existing taxonomy provided by DMOZ. DMOZ is often referred to as the "yellow pages of the internet" and has been extensively used for categorizing webpages and facilitating searching, and also browsing by topics. The DMOZ category tree by itself is very large, with in excess of 500,000 nodes. Therefore it is unusable by itself for ETD categorization. We have pruned the DMOZ category tree and also have enhanced it suitably by making use of taxonomies such as those provided by Proquest, in order to make it more suitable for categorization of ETDs.

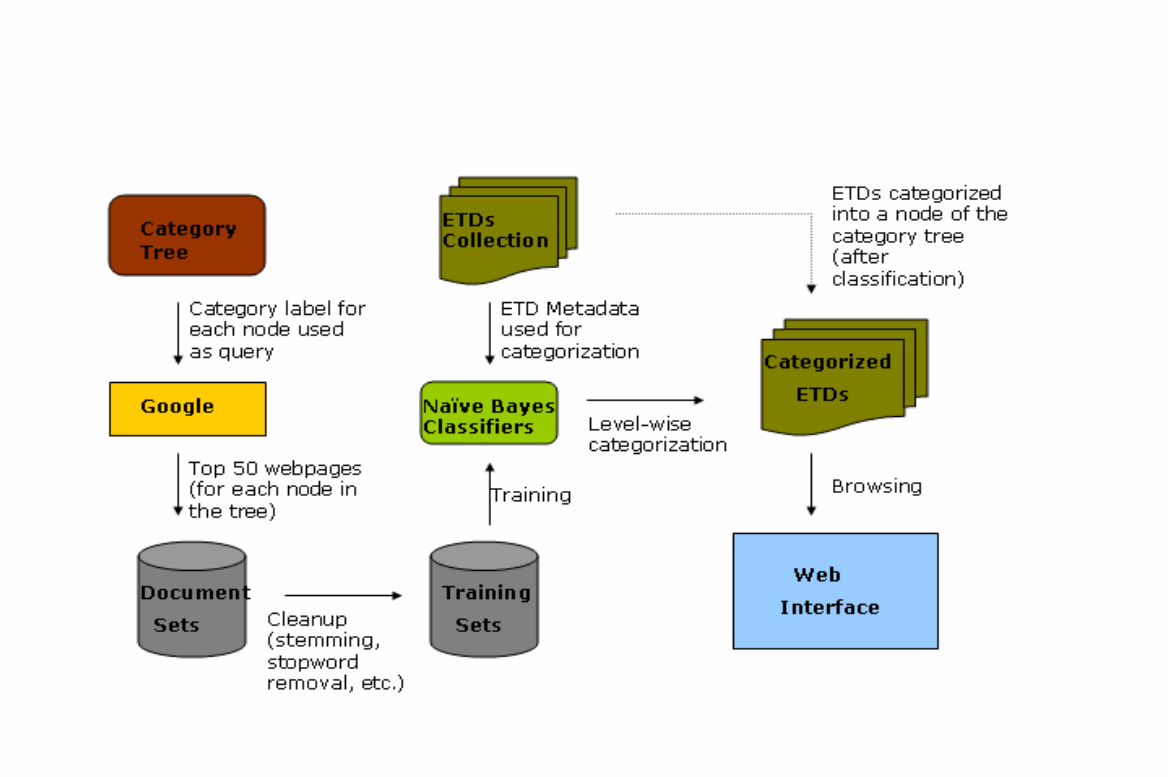

Figure 2. ETD categorization approach

3.3 Crawling ETDs

The Union Catalog provides Dublin Core[6] metadata for over 700,000 ETDs. We have harvested all this information, and stored it locally. We have also made use of the URLs in the metadata field to crawl the ETD itself from the respective university's website, wherever permissible. In this paper however, we make use of only the Dublin Core metadata information to do the categorization. The actual ETDs by themselves will be accessible once we have the web interface ready.

3.3 Categorizing ETDs

The next step is to assign ETDs to their respective topic (or node) in the category tree. We explored various machine learning based document categorization algorithms [8], and decided to suitably modify and use the supervised classification algorithm proposed by Koller et. al.[7]. The detailed steps are as follows:

3.3.1 Building the training set

Given the taxonomy, we first build a collection of documents for each node. We build this document set by using the category label of the node as a query to Google and retrieving the first 50 hits. We then crawl the corresponding webpages, remove the HTML tags, etc. This document set will be used to build the training set for the Naïve Bayes classifiers after some necessary pre-processing.

3.3.2 Training the classifier

We train a Naïve Bayes classifier for each of the nodes in the category tree to distinguish between its children, if any. For example, in Figure 1, we train a Naïve Bayes classifier for the "Top" node so as to be able to assign an ETD to one of the child categories. There will thus be one Naïve Bayes classifier corresponding to each node that has at least 1 child node. Before doing the actual training however, it is necessary to pre-process the document set obtained above. Each document is subject to stopword removal and stemming, and once this is done, the stemmed words are used as features and the Naïve Bayes classifier is trained to distinguish between different child categories.

3.3.3 Categorization

Once the training has been done, the classifiers are used to map ETDs to their respective topic in the tree. We make use of only the metadata information associated with each ETD, viz title, subject, and description fields of Dublin Core, to categorize it suitably. The categorization is done in a level-wise manner. At every level in the category tree only 1 classification task is done. In Figure 1 for example, the first categorization task is done at "Top" to determine to which of the child categories the ETD belongs to. Once this is determined, the process is repeated at the node that was deemed to be the appropriate category in the previous step. This process is continued until the ETD reaches its most specific node in the category tree.

4. ETDs Categorization

The taxonomy that we are developing is based on the DMOZ category tree, and has been pruned and enhanced by using the Proquest taxonomy. The taxonomy currently has 180 nodes, and is 3 levels deep. Currently, we are working on adding lower level nodes. We now present details and results from our categorization experiments.

For our pilot categorization experiment, we selected ETDs from 8 different universities. We have categorized them into the second level of the category tree (as seen in Figure 1), and are working on doing categorization at the lower (more specific) levels in the category tree. Some results relating to this are presented in Table 1.

Since we are doing the categorization into a tree that is only 2 levels deep, we had to build only 1 Naïve Baye's classifier viz. for assigning ETDs into one of the 6 major areas as shown in Figure 1. Training of the classifiers is done offline, and is quite efficient. Training on 300 documents (50 documents for each of the 6 categories) took less than 5 minutes, on a desktop computer with

Dual Core Intel 2.80GHz processors with 1GB memory, and running Ubuntu Linux. Categorization is also very efficient, and to categorize ~74,000 ETDs (Table 1), it took less than 30 minutes.

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**

Table 1. ETD categorization results

The web interface is currently under development, and will be available shortly at http://quantum.dlib.vt.edu/etd/. Our goal with this interface is not just to provide a topical browsing tool, but to enhance it by providing other features like search, browsing by keywords, etc. Once we develop this infrastructure, we will also use it to conduct user studies, in order to measure the quality of the taxonomy, the effectiveness of the categorization algorithm, and the overall user experience.

5. Discussion

In this paper we have presented our pilot work relating to harvesting/crawling and providing suitable access means for large collections of ETDs. Using NDLTD's Union Catalog as the starting point, we have downloaded ETDs from many universities, developed a suitable taxonomy, and used it to categorize the ETDs into topical areas. Even though only the Dublin Core metadata fields have been used for doing the categorization, the ETDs themselves will be available for browsing once the web interface has been completed. We are also rapidly expanding our ETD collection by crawling more dissertations, and categorizing them. Our current focus includes developing a suitable web interface to facilitate browsing access to this collection.

An important future work is to increase the coverage of available ETDs to those beyond the ones in the Union Catalog. We have found this task to be particularly challenging. Many universities restrict access to their ETD collections, or have collections in such a way that they are not amenable to being harvested via automated means. We are making efforts to work with universities to gain access, and to make accessible their ETD collection in a mutually agreeable fashion.

6. References

[1] E. Fox, J. Eaton, G. McMillan, N. Kipp, P. Mather, T. McGonigle, W. Schweiker, and B. DeVane. Networked Digital Library of Theses and Dissertations: An International Effort Unlocking University Resources. D-Lib Magazine, 3(8), Sep. 1997, http://www.dlib.org/dlib/september97/theses/09fox.html

[2] VTLS Visualizer, http://rogers.vtls.com:6080/visualizer, as in May 2009

[3] Scirus ETD Search, http://www.ndltd.org/serviceproviders/scirus-etd-search, as in May 2009

[4] Mozilla Directory, http://www.dmoz.org/, as in May 2009

[5] Proquest Subject Categories, http://www.proquest.com/en-US/products/brands/pl_umi.shtml , as in May 2009

[6] Dublin Core Metadata Initiative, http://dublincore.org/ , as in May 2009

[7] D. Koller and M. Sahami. Hierarchically Classifying Documents Using Very Few Words. In Proceedings of the Fourteenth international Conference on Machine Learning, July 1997, pp 170-178, Morgan Kaufmann Publishers, San Francisco, CA.

[8] F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, Vol 34 No.1, pp 1-47, 2002.