Opening Books and the National Corpus of Graduate Research

William A. Ingram (VT Library), Edward A. Fox (VT CS), and Jian Wu (ODU CS)

Virginia Tech University Libraries, in partnership with Virginia Tech Dept. of Computer Science and Old Dominion University Dept. of Computer Science, requests $506,184 in grant funding for a three-year project to bring computational access to book-length documents, demonstrating that with Electronic Theses and Dissertations (ETDs). Our team has extensive experience in digital libraries, information retrieval, document analysis, and related methods in machine learning and deep learning. We will enhance libraries' ETD programs, devising effective and efficient methods, opening the knowledge currently hidden in this rich corpus of graduate research and scholarship. Our research aligns with (1) the IMLS Strategic Plan to increase public access to important content in a more readable and accessible manner and (2) the National Digital Infrastructures and Initiatives focal area of expanding digital cultural heritage capacities by enabling computational use of collections.

Statement of Need

A fundamental problem in producing the national platform is providing access to book-length material. Despite the growing number of digital books available online, most theoretical and applied research (in text retrieval, extraction, categorization, and summarization) has focused on shorter documents. Discovery and access to scholarly material usually comes through online library catalogs and discovery services, abstracting and indexing databases, and increasingly through large academic search engines like Google Scholar. However, the conceptual models, search algorithms, and techniques underlying such systems are not well-adapted to book-length documents. We need research that focuses specifically on extracting and analyzing segments of long documents (chapters, reference lists, tables, figures), as well as methods for summarizing individual chapters. Research into large collections of books is limited by rights concerns, yet little attention has been given to the very large, analogous collection of millions of ETDs that university libraries house in their institutional repositories.

Most libraries use repository software like DSpace, Samvera/Fedora, or DigitalCommons to house their ETD collections. Metadata is often limited to (some of) the basic Dublin Core elements. These repositories only provide simple tools for accessing whole works, not their components. Full-text indexes, if available, do not allow searching for each of the tables, figures, images, or other data elements contained. Libraries lack the requisite local infrastructure for content mining of their ETDs; their collections are not "computationally amenable." More broadly, nationwide open access services for ETDs generally function at the metadata level, requiring users and content mining systems to go to each repository for download and analysis. Much important knowledge and scientific data lies hidden in ETDs; we need better tools to mine the content and facilitate the identification, discovery, and reuse of these important components.

ETDs are generally made up of multiple chapters and sections but search methods ignore this organization, and ignore requirements for access at different levels of granularity to partly stand-alone entities. For example, most ETDs contain a literature review chapter or a section that reviews prior work. These could be of enormous value to scholars and learners, particularly in specialized fields where good review articles are rare. Further, having an automatically generated summary for each chapter would address the need to speed up discovery of discussions of methods, results, open problems, and other distinctive elements. Many ETDs contain extensive bibliographies, but the citation styles vary widely, making automatic extraction and parsing difficult; existing state of the art methods are tuned, however, to short works in particular domains.

To understand the needs we surveyed a sample of researchers in key areas such as bibliometrics, computational social science, computer science, digital libraries, information retrieval, machine learning, natural language processing, scholarly communication, science of science, web search and mining, and science policy. 18 responded, 74% indicating mining ETDs is valuable for their research communities, 53% indicating they were highly likely to use an ETD infrastructure for their own research, and 68% indicating new collaborations would emerge as a result of this. Respondents also expressed strong interest in data (e.g., citations, figures), software (e.g., knowledge extraction), and research questions (e.g., summarizing long documents).

Project Design

We will answer the following key research questions: (1) *Is it possible to automatically segment ETDs to identify and extract key parts?* (2) *Can we develop effective automatic text summarization techniques for ETD chapters?* (3) *Can we develop more effective access (e.g., search, browse, recommend) techniques that work with ETDs (plus their key parts and summaries), that will support the needs of researchers, scholars, and learners?*

Virginia Tech is an ideal place to lead this research project. VT was the first institution to require ETDs. VT Libraries has one of the largest collections in the US—over 30,000. We will begin in year 1 with those, led by PI Ingram (who directs IT efforts there). Because Co-PI Fox is Executive Director of the Networked Digital Library of Theses and Dissertations, with a union catalog of over 5 million ETD metadata records, we can leverage those, and partner with a representative sample of universities to assemble a research collection of over 200K ETDs by early in year 2. Our deliverables will include results of ETD analysis, shared as they emerge during the project, by identifying, extracting, and devising representations of chapters, reference lists, figures, and tables.

We will prototype infrastructure that can be deployed at individual universities (or consortia) to aid their libraries. This will include an analysis (with text and data mining) system and an access (with searching, browsing, and recommending) system, each prototyped in year 1, refined in year 2, and opened for beta testing early in year 3. PI Ingram will lead the project and focus on summarization, Co-PI Fox on access and summarization, and Co-PI Wu on document analysis. The analysis system will build upon Wu's years of managing the CiteSeerX system at Penn State University, allowing construction of an enormous database of all references in the collection. It will employ new techniques that combine extractive and abstractive methods, including encoder-decoder with attention deep learning models, to for the first time automatically create short summaries for all text chapters of book-length works. The access system will not only work with the metadata and full-text of book-length objects; it will also facilitate access to chapter summaries, chapters, references, figures, and tables.

National Impact

The broad impact of this work will come principally through enhanced access to books and book-length objects that remain woefully underutilized. Additional impact will come from working with a large collection of book-length objects: over 200,000 ETDs. Research universities engage students in research, and much of the graduate level research output is recorded in ETDs. ETDs constitute the majority of the content in university institutional repositories. The ability to understand prior ETDs and leverage them to advance further research and education is in line with university libraries' missions. Facilitating expanded access to, and use of, this important corpus of scholarly content is strongly in the national interest. The ability to computationally process prior ETDs and leverage them to advance further research and education will increase the use of ETD collections and allow libraries to achieve greater impact.

By enhancing the usability of this important body of research, this project has the potential to greatly increase the number of people who benefit from the digital libraries run by universities. It will make ETDs more accessible to a wide spectrum of potential users including but not limited to students, educators, and researchers. The text and data mining process also will reveal unpublished content, which will contribute to the overall fresh knowledge in all academic domains that ETDs cover. The search and recommendation services will provide the most relevant (including unpublished) parts of ETDs to users without requiring them to read entire ETDs, making ETDs more usable, readable, and better cited.

Budget Summary

The estimated total budget is $506,184 over the 3-year period. As lead institution, Virginia Tech requests $212,003 in direct costs and $116,297 in indirect costs. Salary and fringes are being applied to PI Ingram and Co-PI Fox's time ($79,331) and a full-time graduate research assistant ($88,894, plus $43,778 tuition). The subcontract to Old Dominion University ($177,884) will fund Co-PI Wu's salary and fringes and a full-time graduate student.