Scenarios for Advanced Services in an ETD Digital Library

ETD 2017: 20th Int’l Symposium of the NDLTD
Washington, D.C., USA
August 7-9, 2017

Yufeng Ma, Tingting Jiang, Chandani Shrestha, Edward A. Fox
Virginia Tech, Blacksburg, VA 24061 USA
Jian Wu, C. Lee Giles
Pennsylvania State University, University Park, PA 16802 USA

Acknowledgments

• NSF through grant IIS-1423337
• IMLS through grant LG-71-16-0037-16
• Virginia Tech
• Penn State
• NDLTD
Outline

• Introduction
• Advanced Scenarios
  • Table Overview
  • 8 Scenarios
• Key Approaches
  • Building upon Existing Technologies
  • Structured Data Extraction
  • Text Generation
  • Network Visualization
• Conclusion
• References
Introduction

• Vast community of ETD authors
• Even large community who could benefit from ETDs: students, faculty, researchers, scholars, authors, organizations, . . .
• 5 million works in NDLTD Union Catalog
• NDLTD Global Search: faceted search/browse using metadata
• Google Scholar: articles, author profiles, citation data, recommendations based on My Citations, alerts, metrics
• CiteSeerX: documents (summary, citations, active bibliography, co-citation, clustered docs, version history), authors, citation data, tables
• Need for methods and systems for book-size objects
<table>
<thead>
<tr>
<th>Rank</th>
<th>Title</th>
<th>Authors</th>
<th>Year</th>
<th>Source</th>
<th>Language</th>
<th>PubMed ID</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Modeling, design, and optimization of permanent magnet synchronous machines</td>
<td>Angle, Matthew G.</td>
<td>2016</td>
<td>M.I.T. Theses and Dissertations</td>
<td>English</td>
<td></td>
<td>Improvement of performance of robots has necessitated technological advances in control algorithms, mechanical structures, and electric machines. Running, legged robots have presented challenges in the area of electric machinery in particular. In addition to the low-speed, high-torque, low-mass requirements on the machines, the act of running results in an unconventional drive cycle that consists of brief periods of high torque followed by long stretches of minimal torque requirement, a performance envelope that is not matched by...</td>
</tr>
<tr>
<td>2</td>
<td>Long-term, subdermal implantable EEG recording and seizure detection</td>
<td>Do Valle, Bruno Guimaraes</td>
<td>2016</td>
<td>M.I.T. Theses and Dissertations</td>
<td>English</td>
<td></td>
<td>Epilepsy is a common chronic neurological disorder that affects about 1% of the world population. Although electroencephalogram (EEG) has been the chief modality in the diagnosis and treatment of epileptic disorders for more than half a century, long-term recordings (more than a few days) can only be obtained in hospital settings. Many patients, however, have intermittent seizures occurring far less frequent. Patients cannot come into the hospital for weeks on end in order for a seizure to be captured on EEG-a necessary...</td>
</tr>
</tbody>
</table>
| 3    | Parallel algorithms for scheduling data-graph computations           | Hasenplough, William Cleaburn | 2016          | M.I.T. Theses and Dissertations            | English  |           |...
CiteSeerX example
## Advanced Scenarios Utilizing ETDs

<table>
<thead>
<tr>
<th>Stakeholders</th>
<th>Requirements</th>
<th>Expected Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross Cutting</td>
<td>Faceted Browsing</td>
<td>Categorized exploration of ETDs</td>
</tr>
<tr>
<td></td>
<td>Filtered Searching</td>
<td>Metadata-based discovery of ETDs</td>
</tr>
<tr>
<td></td>
<td>Summarization</td>
<td>Synthesis of search results</td>
</tr>
<tr>
<td></td>
<td>Visualization</td>
<td>Linking of related content</td>
</tr>
<tr>
<td>Student Researcher</td>
<td>Aspect-specific access</td>
<td>Specific ETDs, e.g., within a date range or with an advisor name</td>
</tr>
<tr>
<td></td>
<td>Match research question interests</td>
<td>Desired ETDs with quality scores</td>
</tr>
<tr>
<td></td>
<td>Reference extraction</td>
<td>Research questions/hypotheses highlighted</td>
</tr>
<tr>
<td></td>
<td>ETD analysis, Generation of study aids</td>
<td>Related ETDs/books/articles/papers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tabular/Canonical representations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Downloadable package of related work</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lists of journals and conferences</td>
</tr>
<tr>
<td></td>
<td>Linking of problems With methods</td>
<td>ETD content summarizations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Figures, tables, and equations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Key sections and list of related problems</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Visualizations (social/bibliometric networks)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Timeline overview of evolutionary work</td>
</tr>
<tr>
<td>Faculty Researcher</td>
<td>Synthesis of related ETDs</td>
<td>Different methods for a problem</td>
</tr>
<tr>
<td></td>
<td>Proposed approached and solutions</td>
<td>A site with detailed resources</td>
</tr>
<tr>
<td></td>
<td>Future works summarization</td>
<td>An award winning paper (outline/draft)</td>
</tr>
<tr>
<td>Graduate Instructor</td>
<td>Advanced topics, Lecture preparation</td>
<td>Slides cover research question/problems</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Synthesis of provided potential solutions</td>
</tr>
<tr>
<td></td>
<td>Graduate course syllabus formulation</td>
<td>Draft with a hierarchical topical outline</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Link to each topical entry with a reading list</td>
</tr>
<tr>
<td>Conference Organizer</td>
<td>TPC member identification (ID)</td>
<td>List of advisor research faculty names</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ranking tables of advisors</td>
</tr>
<tr>
<td></td>
<td>Potential participants ID</td>
<td>Subgraph of the ETD-derived citation graph</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CSV file of author names, contact info</td>
</tr>
<tr>
<td>Journal Editor</td>
<td>Peer-reviewer ID</td>
<td>Research interest-based reviewer list</td>
</tr>
<tr>
<td></td>
<td>Content originality check</td>
<td>Previous publications of the authors</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Estimated percentage of new content/work</td>
</tr>
</tbody>
</table>
Scenario 1: Identify a reading list

• NS = new graduate student: vague interests, unclear research questions
• NS searches and finds ETDs of interest
• ETDseer produces
  • table (works, quality measures)
  • clusters (each a group of related research questions)
• NS selects some, then receives:
  • Reading list: relevant references (in canonical form)
  • Supplement: figures, tables, equations
• Optional: social/bibliographic networks
Scenario 2: Collect approaches to a research problem

• SR = student researcher
• SR identifies challenging research problem
• SR finds: 3 different approaches, but lacks details, comparisons
• ETDseer: extends the analysis; for each approach produces:
  • Summarization table
  • List of ETDs related
  • List of Source code sites
  • List of Datasets: including training and testing data
  • Values useful for comparison (e.g., time period, # publications)
Scenario 3: Create award-winning paper template

- SR = student researcher
- Almost completed ETD
- SR wants to win best paper award at prestigious conference
- ETDseer does deep analysis of prior award winning papers (& ETDs)
- ETDseer produces a skeleton for the desired paper, from SR’s ETD
  - Detailed outline
  - List of tables, List of figures
  - Equations
  - References
Scenario 4: Identify collaborators

- FR = faculty researcher
- FR describes a research problem requiring collaboration
- ETDseer produces a summary along with:
  - List of ETDs selected (related to research problem)
  - List of documents in their related work sections
  - List of approaches/solutions in the middle of those ETDs
  - List of open problems in ETD conclusion or future work sections
- FR gives feedback: preferences, priorities
- ETDseer produces summary, with shortlist of potential collaborators
  - Contact info, Brief bio-sketches, Notes (complementing FR’s background)
Scenario 5: ETD quality evaluation

• UA: university administrator
• UA seeks assessment of the quality of a locally submitted ETD
• ETDseer produces a report from that ETD:
  • Counts of elements (references, equations, figures, tables)
  • Histogram of citations to key prior works of the author
  • Degree of match between the research problem and the proposed method
  • Summary of experimental results
Scenario 6: Prepare course syllabus and lecture slides

- GI: graduate instructor, teaching a new advanced course
- GI prepares course related materials on a specific research topic
- ETDseer responds with a list of related ETDs
- ETDseer constructs a draft course syllabus:
  - Using clustering, topic analysis, summarization
  - Includes hierarchical topical outline + summaries for each entry
  - Includes reading list = ETDs + open source pubs cited in ETDs
- GI describes a specific problem for course focus
- ETDseer produces list of related ETDs, categorized according to:
  - Problem statements, Research questions, Solutions provided
- Finally, ETDseer produces slides, lecture notes with:
  - Examples, Illustrations, Summary tables
Scenario 7: Organize a conference

- CO: conference organizer
- CO prepares a list of topics from the conference announcement
- ETDseer produces a candidate list of potential members of technical PC
  - Identifies related ETDs, Extracts advisors of authors of those ETDs
  - Extracts authors of highly cited ETDs (at least 5 years old)
  - Ranks on h-index, citation counts, ETD weight in research group
- CO prepares a list of keywords related to the conference theme
- ETDseer produces a list of potential conference authors, attendees
  - Identifies related ETDs, Builds citation graph, Extracts authors
Scenario 8: Manage a journal

• JE: journal editor, seeking reviewers for a paper submission
• JE constructs a query using keywords from the submission
• ETDseer produces a list of researchers with related interests
  • It considers their ETDs + their recent publications
• ETDseer produces a report to aid JE in checking the submission:
  • At least 30% original content relative to author’s prior works
  • Originality relative to works of the identified related researchers
  • Acceptability according to cloud plagiarism detection service
## Advanced Scenarios Utilizing ETDs - top

<table>
<thead>
<tr>
<th>Stakeholders</th>
<th>Requirements</th>
<th>Expected Outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cross Cutting</strong></td>
<td>Faceted Browsing</td>
<td>Categorized exploration of ETDs</td>
</tr>
<tr>
<td></td>
<td>Filtered Searching</td>
<td>Metadata-based discovery of ETDs</td>
</tr>
<tr>
<td></td>
<td>Summarization</td>
<td>Synthesis of search results</td>
</tr>
<tr>
<td></td>
<td>Visualization</td>
<td>Linking of related content</td>
</tr>
<tr>
<td><strong>Student</strong></td>
<td>Aspect-specific access</td>
<td>Specific ETDs, e.g., within a date range or with an advisor name</td>
</tr>
<tr>
<td></td>
<td>Match research question interests</td>
<td>Desired ETDs with quality scores</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Research questions/hypotheses highlighted</td>
</tr>
<tr>
<td></td>
<td>Reference extraction</td>
<td>Related ETDs/books/articles/papers</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tabular/Canonical representations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Downloadable package of related work</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lists of journals and conferences</td>
</tr>
<tr>
<td><strong>Researcher</strong></td>
<td>ETD analysis, Generation of study aids</td>
<td>ETD content summarizations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Figures, tables, and equations</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Key sections and list of related problems</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Visualizations (social/bibliometric networks)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Timeline overview of evolutionary work</td>
</tr>
<tr>
<td></td>
<td>Linking of problems With methods</td>
<td>Different methods for a problem</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A site with detailed resources</td>
</tr>
<tr>
<td></td>
<td></td>
<td>An award winning paper (outline/draft)</td>
</tr>
</tbody>
</table>
## Advanced Scenarios Utilizing ETDs - bottom

<table>
<thead>
<tr>
<th>Stakeholders</th>
<th>Requirements</th>
<th>Expected Outputs</th>
</tr>
</thead>
</table>
| Faculty Researcher| Research problem exploration aid                  | Synthesis of related ETDs
                             | Proposed approached and solutions
                             | Future works summarization                                                      |
| Graduate Instructor| Advanced topics, Lecture preparation              | Slides cover research question/problems
                             | Synthesis of provided potential solutions                                         |
| Conference Organizer| Graduate course syllabus formulation              | Draft with a hierarchical topical outline
                             | Link to each topical entry with a reading list                                   |
| Journal Editor    | TPC member identification (ID)                    | List of advisor research faculty names
                             | Ranking tables of advisors                                                      |
|                   | Potential participants ID                         | Subgraph of the ETD-derived citation graph
                             | CSV file of author names, contact info                                           |
|                   | Peer-reviewer ID                                  | Research interest-based reviewer list                                             |
|                   | Content originality check                         | Previous publications of the authors
                             | Estimated percentage of new content/work                                         |
Key Approaches: Building Upon Existing Technologies

• NDLTD has relevant metadata
• Pilot studies at VT have leveraged that to harvest thousands of PDFs
• CiteSeerX, part of SeerSuite, has demonstrated key services
  • Mostly for CS or Chemistry-related works, with tables, figures
  • Using knowledge bases, heuristics, regular expressions, classifiers
• Challenges beyond CiteSeerX methods:
  • All disciplines, all styles, all writing formats
  • Extracting passages and hard-to-specify text blocks: hypotheses
  • Need for robust, extensible methods
Key Approaches: Structured Data Extraction

• Above-mentioned challenges suggest using deep (machine) learning
• Locating, analyzing, and representing references
  • End of work, end of chapter, end of page
  • Thousands of styles + variations + author improvisation
  • Ambiguities: authors, venues, missing information
• Document segmentation
  • Book-like structure vs. collection of published papers
  • Inconsistencies: front-matter vs. body, idiosyncratic taxonomies
  • Tables and figures: domain-specific conventions, author ingenuity
Key Approaches: Text Generation

• Leveraging segmentation for the following:
• Passage retrieval leveraging discourse and semantic analysis
• Sub-languages: different jargon; by (sub)discipline, multi-discipline
• Salient keywords: overall, per segment/unit
• Extracting multiple topics
  • Over: group of ETDs, one ETD, one or multiple chapters or sections
  • LDA, Word2vec (for sub-language), encoder-decoder, attention, ...
Key Approaches: Network Visualization

- ETD-ETD, ETD-Paper, Paper-Paper, Author-Document, Author-Author
- Attributes: Citation counts, Paper quality, Author publication count
- Relationships: Student-Advisor, Co-advised, Author-of, Cites, Co-cited
- Extended relationships: Co-authors, Panelists, Related research

- Force-directed graph visualization
- Path-based queries of networked information
Conclusion

• Now have basic services using metadata
• Need methods that can leverage full-text

• Scenarios for each of diverse set of stakeholders

• Leverage NDLTD, CiteSeerX, Deep learning

• Broad program of research needed, leading first to useful prototypes

• Broad impact on research, education, scholarly activities
References - 1

• Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent dirichlet allocation."
• Choudhury, Sagnik Ray, Prasenjit Mitra, Andi Kirk, Silvia Szep, Donald Pellegrino, Sue Jones, and C Lee Giles. 2013. "Figure metadata extraction from digital documents."
• Choudhury, Sagnik Ray, Shuting Wang, and C Lee Giles. 2016b. "Scalable algorithms for scholarly figure mining and semantics."
• He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. "Mask r-cnn."
• Liu, Xiaoyong, and W Bruce Croft. 2002. "Passage retrieval based on language models."
• Liu, Ying, Kun Bai, Prasenjit Mitra, and C Lee Giles. 2007. "TableSeer: automatic table metadata extraction and searching in digital libraries."
References - 2

• Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. "Distributed representations of words and phrases and their compositionality."


• Ororbia II, Alexander G, Jian Wu, Madian Khabsa, Kyle Williams, and Clyde Lee Giles. 2015. "Big scholarly data in CiteSeerX: Information extraction from the web."

• Pennington, Jeffrey, Richard Socher, and Christopher D Manning. 2014. "Glove: Global vectors for word representation."

• Ray Choudhury, Sagnik, and Clyde Lee Giles. 2015. "An architecture for information extraction from figures in digital libraries."

• Ray Choudhury, Sagnik, Prasenjit Mitra, and Clyde Lee Giles. 2015. "Automatic extraction of figures from scholarly documents."


• Srinivasan, Venkat, Mohamed Magdy, and Edward A Fox. 2011. "Enhanced browsing system for Electronic Theses and Dissertations."

• Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. 2014. "Sequence to sequence learning with neural networks."


• Wade, Courtney, and James Allan. 2005. Passage retrieval and evaluation.

Questions?
Discussion?
Recommendations?

Thank You!

fox@vt.edu
http://fox.cs.vt.edu