ETDseer
Final Term Term Project Presentation

Yufeng Ma, Tingting Jiang, Chandani Shrestha
CS 6604 - Digital Libraries
Virginia Tech, Blacksburg, VA, 24061
Professor Edward A. Fox
April 27, 2017
Overview

● Problem Statements/Motivation
● Related Works
  ● NDLTD
  ● CiteSeer
● Stakeholder Overview
  ● Architecture
  ● Scenarios and Services
    ○ Workflow example

● Structured Data Extraction
  ● ETD Segmentation
  ● Table & Figure Extraction
  ● Reference Extraction
● Text Summarization
● Network Visualization

● Working On...
Background

**Problem Statements**
- ETDs as international resource - extensive potential
- Largely Untapped
- Limitation in existing related tools
  - Document length, Document accessibility (Full Text)
  - Summarization, Visualization
- Challenges working with ETDs
- Need for enhanced usability

**Motivation**
- Rich knowledge base - Single platform
- Accessible to broader group of users
- Network of institutional repositories
Related Works

- NDLTD
- VTechworks
- CiteSeer
- ContentMine
NDLTD

- ETDs Statistics
- Functional Limitations
  - Categorization Missing
  - No Full Text Access
Data: Make use of Metadata

Technologies: SeerSuite tools
  - Automatic citation indexing
  - Automatic metadata extraction
  - Reference linking
  - Author disambiguation
  - Related documents
  - Full-text indexing
Stakeholders Overview

Student Researcher

Conference Organizer

Faculty Researcher

ETDSeer

Journal Editor

Graduate Instructor
Architecture
## Scenario 1 - Student Researcher

<table>
<thead>
<tr>
<th>Requirements</th>
<th>Key techniques</th>
<th>Expected Outputs</th>
</tr>
</thead>
</table>
| Metadata-based search                          | SeerSuite         | • Specific ETDs within a date range  
• Specific ETDs with an advisor name          |
| Research interests discovery                   | Structured text extraction | • Desired ETDs with quality scores  
• Research questions/hypotheses highlighted   |
| Reference extraction                           | Structured text extraction | • Related ETDs/books/articles/papers  
• Tabular/Canonical representations  
• Downloadable package of related work  
• Lists of journals/conferences             |
| Linking of problems with methods              | Text extraction   | • Different methods for a problem  
• A site with detailed resources  
• An award winning paper (outline/draft)      |
| ETD analysis and study aids                   | Deep learning     | • ETD content summarizations  
• Figures, tables and equations  
• Key sections and list of related problems  
• Visualizations (social/bibliometrics networks)  
• Timeline overview of evolutionary work      |
An Example Workflow
## Other Scenarios

<table>
<thead>
<tr>
<th>Stakeholder</th>
<th>Requirements</th>
<th>Expected Outputs</th>
</tr>
</thead>
</table>
| Faculty Researcher      | Research problem exploration aid      | ● Synthesis of related ETDs  
                          |                                                                                   | ● Proposed approaches/solutions  
                          |                                                                                   | ● Future works summarization     |
| Graduate Instructor     | Graduate course syllabus formulation  | ● Draft with a hierarchical topical outline  
                          |                                                                                   | ● Link to each topical entry with a reading list |
|                         | Advanced topic, lecture preparation   | ● Slides cover research questions/problems  
<pre><code>                      |                                                                                   | ● Synthesis of provided potential solutions |
</code></pre>
<table>
<thead>
<tr>
<th>Stakeholder</th>
<th>Requirements</th>
<th>Expected Outputs</th>
</tr>
</thead>
</table>
| Conference Organizer | TPC member identification                       | • List of advisor research faculty names  
                              |                                                                                  | • Ranking table of advisors                                                   |
|                      | Potential participants identification            | • Subgraph of ETD-derived citation graph  
                              |                                                                                  | • CSV file of author names, contact info.                                     |
| Journal Editor       | Peer-reviewer identification                    | Research interest-based reviewer list                                              |
|                      | Content originality check                       | • Previous publications of the authors  
                              |                                                                                  | • Estimated percentage of the new content/work                                |
Structured Data Extraction:
ETD Segmentation

- Heuristics-based strategy
  - Start with ‘Chapter’ or ‘CHAPTER’
  - Font size and style

- Deep learning approaches
  - Treat every two pages as one image
  - Manually label each image - chapter breaking point or not
  - Build a CNN model for classification
Structured Data Extraction: Table & Figure Extraction

- Extend work from TableSeer
  - Table box detection
  - Table metadata extraction
  - Deal with styles of more disciplines

- Figure Extraction
  - Sagnik Ray Choudhury’s work
  - CNN based approach - Mask R-CNN
  - AMT for figure labeling

* Y. Liu, et al., TableSeer, JCDL 2007.
* S. R. Choudhury, et al., An Architecture for information extraction from figures in DLs, WWW, 2015.

* Y. Liu, et al., TableSeer, JCDL 2007.
* S. R. Choudhury, et al., An Architecture for information extraction from figures in DLs, WWW, 2015.
Structured Data Extraction: Reference Extraction

- References that appear at the end
  - SeerSuite

- References that appear anywhere like footnotes
  - Deep learning for learning reference features
  - Classifier will be trained

- Represented in canonical format like BibTeX

---

Donald E. Knuth’s Principal Principle is one of my favourite principles. One of my favourite books is *TeXbook*. Everybody should be rational. Knuth said a lot of things. For instance, he said that everybody should be rational, and he said that everybody should drive on the right side of the road. Arnold van Gennep said that everybody should drive on the left, but otherwise van Gennep’s work agrees with Knuth.

---


5Van Gennep, *Rites de passage*, p. 4.
Text Summarization

- **Topic modeling based approach**
  - Extracted keyword or phrase
  - Probabilistic graphical model

- **Deep learning approach**
  - Complete sentence
  - Attention model
  - Pointer & Generator
Network Visualization

- **Reference Network**
  - Given one ETD, citation relationship between papers and ETDs is visualized

- **Social Network**
  - Collaboration strength between research groups
  - Author social network
Working on.....

- User Study- based on the prospective Stakeholders
  - Questionnaires to reach broader audience
  - Interview (Focus group)
  - Hands on use and feedback

- Analysis of the collected data

- Presentation/Discussion of the result based on the analysis


Complete list of references: https://goo.gl/QVgPQN
Thank you!!!