CS 6604 Digital Libraries

Final Term Project Report
ETDseer: Concept Paper

Team Members:
Yufeng Ma
Tingting Jiang
Chandani Shrestha

5/3/2017

Virginia Tech
Blacksburg, VA 24061

Instructor: Professor Edward A. Fox

Draft Project Summary

ETDseer (electronic thesis and dissertation digital library connected with SeerSuite) will build on 15 years of collaboration between teams at Virginia Tech (VT) and Penn State University (PSU), which both have been leaders in the worldwide digital library community. VT helped launch the national and international efforts for ETDs more than 20 years ago, which has been led by the Networked Digital Library of Theses and Dissertations (NDLTD, directed by PI Fox); its Union Catalog has increased to 5 million records. PSU hosts CiteSeerX, which co-PI Giles launched almost 20 years ago, and which is connected with a wide variety of research results under the SeerSuite family.

ETDs, typically in PDF, are a largely untapped international resource. Digital libraries with advanced services can effectively address the broad needs to discover and utilize ETDs of interest. Our research will leverage SeerSuite methods that have been applied mostly to short documents, plus a variety of exploratory studies at VT, and will yield a web of graduate research, rich knowledge bases, and a digital library with effective interfaces. References will be analyzed and converted to canonical forms, figures and tables will be recognized and re-represented for flexible searching, small sections (acknowledgments, biographical sketches) will be mined, and aids for researchers will be built, especially from literature reviews and discussions of future work. Entity recognition and disambiguation will facilitate flexible use of a large graph of linked open data.

Keywords: CiteSeerX, deep learning, domain independent digital library (DL), information extraction (IE), information retrieval (IR), machine learning (ML), natural language processing (NLP), NDLTD

Intellectual Merit
This research will lead to fundamental contributions in the DL, IE, IR, ML, and Web crawling/archiving areas, as well as contributions in HCI and AI. For example, we will develop and apply new methods to find all references (from footnotes, ends of chapters, and the bibliography), extract unambiguous values in spite of thousands of different reference styles employed, and build end-user services atop an accurate citation graph including distinct works as well as entities for graduate students, advisors, and other scholars.

Broader Impacts
This research will tackle many difficult challenges faced by the AI, DL, IE, IR, ML, and NLP communities, at scale, advancing big data efforts in these domains. It will have broad impact on national research and education by opening up ETD resources for broad and helpful use.

PROJECT DESCRIPTION

Using advanced IE and NLP techniques, coupled with machine learning and information retrieval methods, we will research ETDseer, a tailored DL for large English ETD collections, which would offer special services: review the literature, identify hypotheses, list research questions, explain approaches, describe methods, summarize results, discuss findings, present conclusions, and provide insights on open problems. We can provide these services by processing references and citations, as well as information extracted from chapters, sections/subsections, tables, and figures. Though ETD collections cover many disciplines, a suitable domain independent digital library can be prototyped now, using advanced natural language processing and information extraction techniques, coupled with machine learning and information retrieval methods. The resulting system would enable stakeholders to engage in advanced scenarios that go well beyond conventional searching and browsing. Section 1 summarizes the statement of need for this research with goals and objectives. Section 2 covers results from prior NSF support. Section 3 describes our approach, including connections with real life scenarios for the scholarly community, methodologies applied, and system components. Section 4 addresses project management and timing. Remaining sections give Related Work, Expected Significance, and Broader Impacts.

1 Introduction

Statement of Need

ETDs can be a valuable aid to learning and scholarship. They are a largely unused international resource.Tapping this resource requires research first on effective ways to discover ETDs of interest, and then support for better using identified works to aid a highly diverse community of researchers and educators. At present, there is no good way to address these needs for large heterogeneous collections of long documents like ETDs, typically in PDF, but also available in other formats.

Though there are widely available DLs that support indexing, searching, and browsing of short articles or papers, book length objects typically are searched using limited and often misleading metadata records, with minimal help from classification efforts (especially for multidisciplinary works). While full-text search is used to extend faceted searching, it is not effective when processing large files in natural language. A preferable approach could be applying Natural Language Processing (NLP), e.g., information extraction (IE). Currently available resources like Google Scholar and CiteSeerX [6, 20, 21, 30, 39, 54], that extract, analyze, and link references in short publications, do provide additional capabilities, but they do not work well with ETDs (due to length, complexity, and domain variations).

In this research, we address the need for a tailored DL focusing on ETDs based on advanced concepts from information retrieval (IR), Web crawling/archiving, information extraction, machine learning (ML), and deep learning.

Goals

1. Aid effective dissemination of valuable information from years of collections of ETDs, to the scholarly community, including a whole spectrum of researchers from an inexperienced student researcher to a well-established faculty member.

2. Prototype a domain independent digital library that provides advanced services to a vast collection of book-length long documents.

3. Advance the state-of-the-art in DL and NLP with regard to long document retrieval and utilization, for both individual documents and groups of documents.

4. Advance the state-of-the-art in DL and NLP with regard to analyzing long documents; processing tables, figures, and references; extracting citations and key concepts from chapters, sections, and subsections (including special parts of ETDs like literature reviews, acknowledgements, and bio-sketches); and expanding the capabilities of DL beyond indexing, searching, and browsing.

5. Advance the state-of-the-art in information integration, synthesis, and generation, including when generating lists (e.g., of references, literature reviews, approaches, methods, and findings).

Objectives

1. Develop methods to extract semantic content and surrounding context from long documents such as research questions and their corresponding hypotheses.

2. Aid stakeholders through interfaces tailored for exploration, to discover interesting content, review literature, and evaluate references as well as ETDs.

3. Assist stakeholders as they study and research with ETDs by intelligently providing advanced services through stating hypotheses, listing research questions, explaining approach, describing methods, summarizing results, discussing findings, drawing conclusions, and providing insights about open problems.

2 Approach

2.1 Advanced services in discovering and utilizing ETDs

ETDseer is aiming to improve the utilization of valuable ETD resources by providing specialized services to a spectrum of scholarly community users in a fine-grained manner. Based on the specific research and study needs of different stakeholders, we identify some of the representative real-life scenarios and describe them as follows. Table 1 summarizes the service requirements generated from the envisioned scenarios and their corresponding anticipated outputs.

Scenario 1. Identify a reading list
NS, a new graduate student who enters the research arena with vague interests, needs to study published works to gain understanding of key research questions. NS can search

and find a suitable set of ETDs. Results are given in a tabular form, indicating the quality of the selected works (based on citation counts and other criteria), and showing clusters of related research questions. After NS reviews this data and selects portions of particular interest, the DL extends its analysis. Relevant references are extracted, converted to a canonical form, and presented as a reading list. Additionally, the figures, tables, and equations of the selected ETDs are summarized and presented as a supplement. Optionally, social/bibliographic networks, and other helpful visualizations, are provided.

Scenario 2. Collect approaches to a research problem
SR, a student researcher, has come across a challenging research problem and is interested in the discussions in journal and conference papers he has reviewed so far, which indicate that three different methods have been employed, but without details and comparative studies. ETDseer helps SR identify the ETDs that are related to each of the methods as well as corresponding involved datasets. Then ETDseer creates a site that has descriptions of the source code, as well as the training and testing data associated with each method. A well-formatted summarization table is generated.

Scenario 3. Create award-winning paper template
Student researcher, SR, with an almost completed ETD, wants to win the best paper award at a prestigious conference. Based on deep learning analysis of other award winning papers in that area, and their corresponding ETDs, an outline of a paper derived from the ETD is constructed for SR, including tables, figures, equations, and references.

Scenario 4. Identify Collaborators
Faculty researcher, FR, who has identified a specific research problem that necessitates collaboration, seeks a list of different approaches used to tackle this problem as well as a timeline view of the evolution of associated research studies. FR describes the problem, and receives a list of selected ETDs. Documents listed in the related work sections, proposed approaches/solutions in the middle of ETDs, and open problems mentioned in the conclusion or the future work sections, are identified. A summary table categorizing the details is presented.

FR studies the summary provided and provides feedback about preferences and priorities. The DL prepares a tailored summary, and a shortlist of potential collaborators, along with their contact information and brief bio-sketches, that is supplemented with notes on how they might complement FR's background.

Scenario 5. ETD quality evaluation
University administrator, UA, would like a rough assessment of the quality of an ETD submitted from one of the local departments. ETDseer provides a table related to the selected ETD that contains: counts of elements (references, equations, figures, and tables), a histogram of citations to key prior works of the author, degree of match between proposed approach and research problem, and a summary of experimental results.

Scenario 6. Prepare course syllabus and lecture slides
Graduate instructor, GI, is teaching a new advanced course. GI prepares course related

materials on a specific research topic and receives a list of related ETDs in ETDseer. Based on the ETDs of interest, ETDseer uses clustering, topic analysis, and summarization methods to construct a draft course syllabus. Included in the syllabus is a hierarchical topical outline, with summaries for each entry, linked with a suitable reading list, which includes the ETDs as well as the most important other open source publications that were discussed in those ETDs.

GI also wants to focus on a specific problem and discuss the most promising solutions. GI gives ETDseer a description of the problem, and receives related ETDs that are neatly categorized in terms of their various problem statements, research questions, and provided solutions. Furthermore, ETDseer creates drafts for class including properly sequenced slides and lecture notes, with helpful examples, illustrations, and summary tables.

Scenario 7. Organize a conference
A conference organizer, CO, wants to identify a list of Technical Program Committee members for a conference. CO gives ETDseer the list of topics from the announcement. ETDseer searches through the related ETDs and returns a list of advisor research faculty names that appear in the metadata for the ETDs, along with names of ETD authors from at least five years earlier who have highly cited ETDs. To provide CO with more detailed information, ETDseer generates a table that ranks those advisors based on h-index, the weight of the ETDs in each advisor's research group, citation counts, etc. CO also wants to identify the potential participants of the conference. CO queries ETDseer with a list of keywords, related to the theme of the conference. ETDseer presents a subgraph from the relevant portion of the ETD-derived citation graph, extracts authors of those works, and returns their names and contact information as a CSV file, which can be used to send a general conference announcement for submissions and/or participation.

Scenario 8. Manage a journal
A journal editor, JE, seeks to identify peer reviewers for a journal paper submission. JE queries ETDseer using keywords from the submission. ETDseer responds with several author names, indicating research interest closely related to the submission. This would be based on their published ETDs and their recent publications that can be extracted by ETDseer. JE also needs to check if a follow up paper submission has at least 30% original content relative to previous publications. JE queries ETDseer with the author names for an originality check. ETDseer identifies previous publications belonging to the authors, and uses a cloud service to return the estimated percentage of new content/work in the submitted paper.

2.2 System Design

We begin our discussion by describing the architectural design of ETDseer that extends CiteSeerX [6] to provide advanced services on the ETD collections that enhance learning and research. An outline of the ETDseer architecture that is based on the 5S framework [23] is shown in Figure 1.

Table 1: ETDseer Scenario Table

Stakeholder Group(s)	Requirements	Expected Outputs
Cross Cutting	Faceted Browsing	Categorized exploration of ETDs
	Filtered Searching	Metadata-based discovery of ETDs
	Summarization	Synthesis of search results
	Visualization	Linking of related content
Student	Aspect-specific access	Specific ETDs, e.g., within a date range or with an advisor name
	Match research question interests	Desired ETDs with quality scores Research questions/hypotheses highlighted
Researcher	Reference extraction	Related ETDs/books/articles/papers Tabular/Canonical representations Downloadable package of related work Lists of journals and conferences
	ETD analysis, Generation of study aids	ETD content summarizations Figures, tables, and equations Key sections and list of related problems Visualizations (social/bibliometrics networks) Timeline overview of evolutionary work
	Linking of problems with methods	Different methods for a problem A site with detailed resources An award winning paper (outline/draft)
Faculty Researcher	Research problem exploration aid	Synthesis of related ETDs Proposed approaches and solutions Future works summarization
Graduate Instructor	Advanced topics lecture preparation	Slides cover research question/ problems Synthesis of provided potential solutions
	Graduate course syllabus formulation	Draft with a hierarchical topical outline Link to each topical entry with a reading list
Conference Organizer	TPC member identification	List of advisor research faculty names Ranking table of advisors
	Potential participants identification	Subgraph of the 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 new content/work

2.2.1 Data Source

ETDseer will be based on ETD submissions to NDLTD [1, 17]. The NDLTD initiative, which was started in the 90s, seeks to expand electronic publication of student research, and make ETDs accessible from around the world. The Union Catalog consists of meta-data records for ETDs from contributing institutions and universities. As of March 2010,

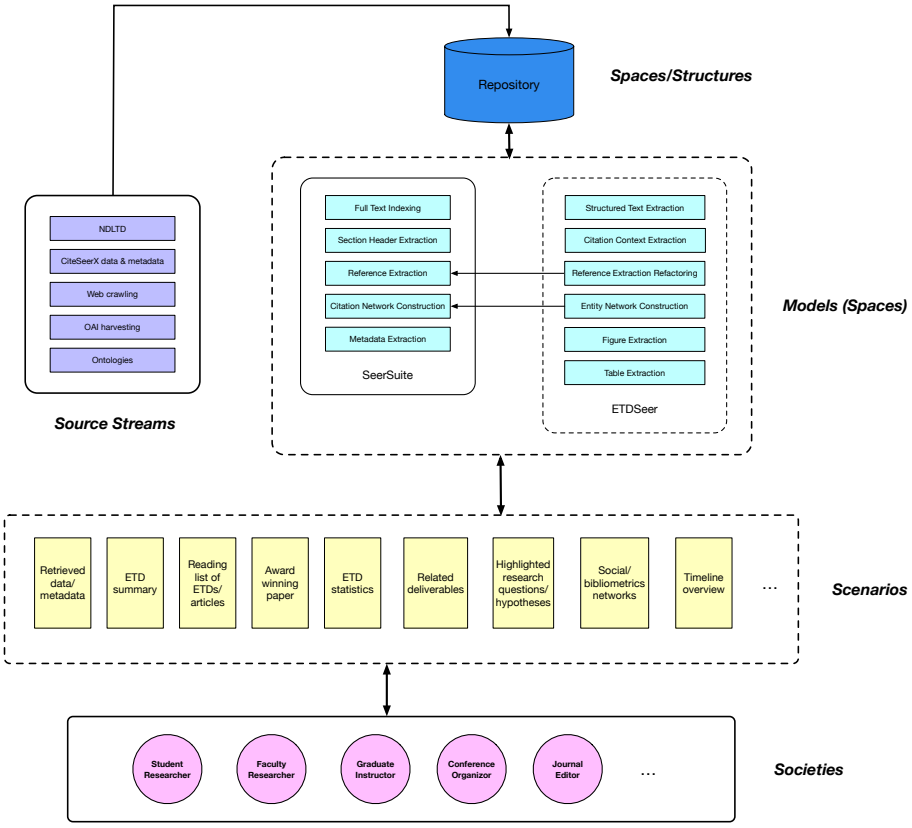

Figure 1: ETDseer System Architecture

the Union Catalog consisted of metadata records for 820,000 ETDs in various languages. By 2013 the number exceeded three million and by April 2017, five million. It contains ETDs in at least 12 languages. It is the single largest cumulative source of information on ETDs available on the Internet. Figure 2 shows the top 12 NDLTD participating members ranking by the number of ETD submissions as of this writing.

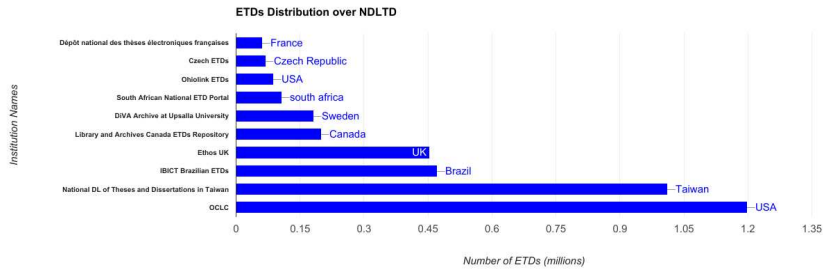

Figure 2: Top 12 NDLTD participating institutions and universities

These book-sized ETDs are global scholarship assets, which often have higher value than shorter scholarly publications (e.g., conference and journal papers) in terms of rich information content and reproducibility of science based on detailed descriptions of research

work. Easier access and practical services enabling full usage of this vast collection of rich documents will be of invaluable aid to the scholarly community.

2.2.2 Existing ETD-related DL Technologies

ETDseer will hold the ability to access, analyze, and synthesize from long ETD documents, with many more advanced services than are available in DLs like NDLTD [1]. NDLTD provides basic functionalities that support collecting, disseminating, and searching/browsing of ETDs. In particular, it provides filtered searches with customizable fields like subject, publication date, and language as well as faceted browsing with language and keyword predefined as the two facet fields. Figure 3 shows the filtered/faceted navigation capabilities in NDLTD. As we can see, it misses quite a few fields such as discipline, research questions/problems, and citations that are crucial to the ETD content space.

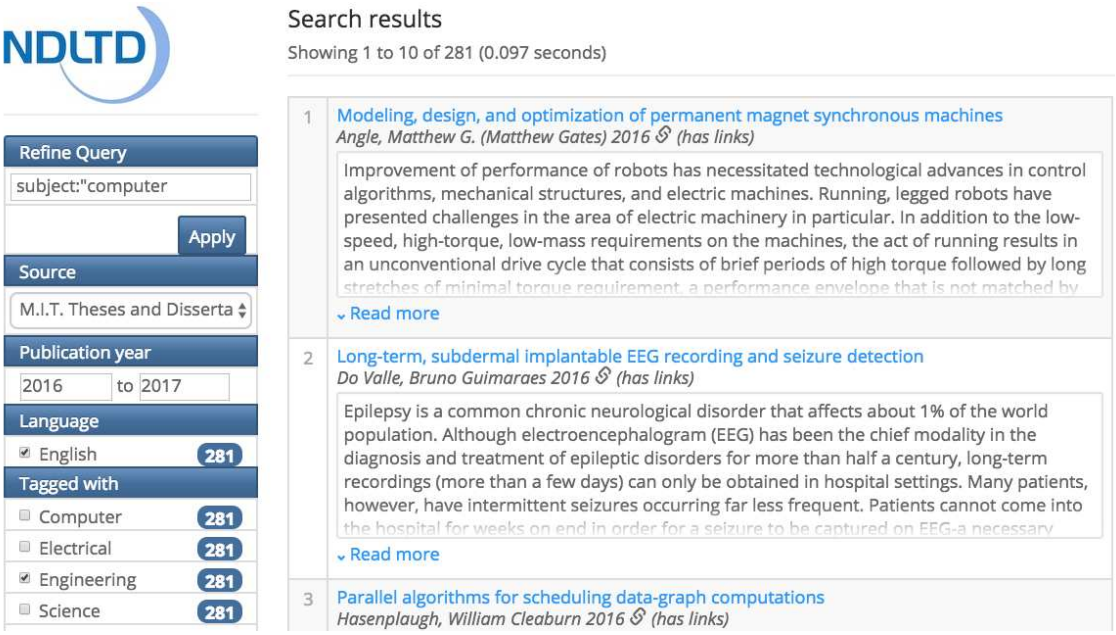

Figure 3: Filtered and faceted navigation in NDLTD

On the other hand, a closely related DL, CiteSeerX [30], does a much better job in terms of analyzing and presenting the scholarly information from large collections of research works, with features like automatic metadata extraction, citation indexing, and reference linking. However, CiteSeerX deals only with short length documents, thus excluding its use of many features in long ETD documents.

A proper integration of these two DLs, where the datasets (metadata and long documents) are mostly from NDLTD, combined with the metadata and technical features from CiteSeerX, will guide the initial development of ETDseer. ETDseer will leverage most of the DL tools that CiteSeerX has provided. Nevertheless a simple integration of features is neither feasible nor sufficient with ETDs. Although CiteSeerX can work as a technological foundation for building ETDseer, the target documents that CiteSeerX serves are short

documents. To process and analyze much longer documents in ETDseer and to achieve similar goals, heuristic approaches such as regular expressions and knowledge rule-based approaches adopted by CiteSeerX will not be sufficient. Furthermore, unlike CiteSeerX, most of the functional requirements in ETDseer, such as extracting research questions and hypotheses, cannot be addressed via keyword and fulltext only searching and browsing.

CiteSeerX handles documents from several specific scientific disciplines, mostly from computer science and a few from chemistry. The writing format and styles of these domain specific documents are fairly consistent, so a knowledge base can be built and used accordingly. However, ETDs can be from multiple disciplines, with various formats and writing styles, which further adds to the complexity of how they are analyzed and presented. Apart from the main content and format, the multidisciplinary ETDs will vary in the construction and frequency of other multimedia aspects like tables, equations, graphs, and images. Additionally, depending on the domain of interest, references can appear in the final section, at the end of chapters, or as footnotes. These unique characteristics of ETDs have to be taken into consideration, and machine learning approaches, such as neural networks-based deep learning methods, can potentially be used to achieve the advanced functionalities of ETDseer.

Besides leveraging existing DL technologies, we will research how to address the broad and deep requirements brought by ETDseer stakeholders along with ETD-related challenges.

2.2.3 Structured Data Extraction

Since ETDs are long documents, a passage retrieval, combined with a document retrieval, approach is appropriate [44, 51, 37]. Furthermore, ETDseer should extract key information from ETDs and present that to end users. However, structural complications associated with ETDs make these retrieval and extraction tasks much more difficult than when dealing with representative scientific papers. In a scientific paper, the authors typically arrange contents into clearly distinct sections, such as introduction, related work, conclusions, and references. In contrast, ETDs have highly varied and unpredictable structures. Some ETDs have book-like structures where each chapter has its designated role in the book, while others are presented as a composite of scientific papers where each chapter represents a separate paper. In the latter case, front matter (i.e., table of contents, table of figures, and table of tables) can appear in multiple chapters. The location of references can vary: some appear at the end of the document, some are at the end of chapters, and some can be seen as footnotes [40]. Existing SeerSuite segmentation and extraction tools are not effective with such varying formats, or with unpredictable structures. Herein we propose potential solutions so ETDseer can address these challenges.

Research Questions

- How can ETDseer identify all hierarchically arranged sections in a long document?
- How can we extract (or infer) from the ETDs the table of figures, table of tables, table of equations, as well as each of the figures, tables, and equations, along with their content/captions?

- For references that appear in various places within the ETDs, how can we locate and extract them, and present them in a canonical format?

Research Plan

- *Task 1:* **Segmentation** of ETD documents can be achieved through the combination of heuristic-based strategies [47] and deep learning approaches. Heuristic methods will focus on ETD specific knowledge like font and semantic information of chapter titles and numberings. To deal with ambiguous chapter segmentations, deep learning approaches will treat the whole document as one picture, with which segmentation or detection approaches like Mask R-CNN [26] can be applied. With the resulting structured data, key extraction methods used in CiteSeerX [6, 21] can be applied to obtain the hierarchically structured subsections from the ETDs.

- *Task 2:* **Tables and figures** are effective devices for presenting research results. To extract these types of structured data, Dr. Giles (Co-PI) and his research group have made good progress on dealing with scientific data in computer science and chemistry related domains. This has led to TableSeer [33] and research results from the doctoral work on figures by Sagnik Ray Choudhury [57], along with other related publications [12, 13, 43, 10, 11].

  With discipline-independent ETDs, the content in figures and tables can be domain specific. Figure 4 shows figures and tables from various disciplines. In Computer Science and Mathematical disciplines they have more data-intensive content, while disciplines such as Education and Sociology tend to have table and figure content with more descriptive words. To extract these style-free figures and tables, machine-learning based approaches should be applied in addition to existing CiteSeerX methods.

- *Task 3:* Existing **citation extraction** tools from CiteSeerX perform well on extracting references at the end of documents and thus they can be directly adopted by ETDseer. For references occurring at the end of each chapter, or some other random locations like placed in tables or as footnotes, we choose to use machine learning-based approaches. In particular, we will first conduct studies of ETDs to gather reference placement information from a wide range of disciplines. This step will ensure that a variety of features will be extracted. Then we can perform fine-grained feature selection to improve accuracy and efficiency. A classifier will be trained to identify references appearing at unusual locations. Extracted references will be ultimately converted into some canonical format like BibTex so they can be neatly presented in any user-desired style.

2.2.4 Text Generation

While NDLTD has made access to ETDs much easier than before, its keyword-based searching and topic-based faceted browsing greatly limit its usage. A large number of ETDs will be returned on a particular topic, but it will be difficult for users to identify which ETDs to read. Besides, each of these book-length documents generally contains

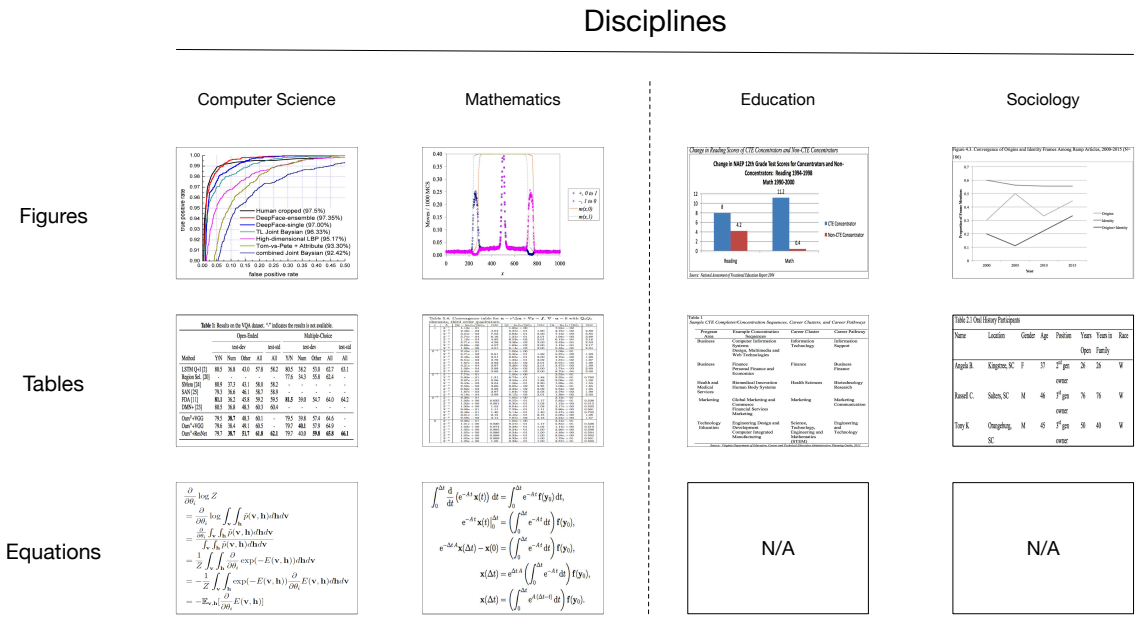

Figure 4: Sample figures, tables and equations in ETDs from various disciplines

numerous chapters, within which there are multiple sections and subsections. It is time consuming for readers to browse through these length documents with their numerous chapters and sections, when only interested in the key ideas. A good summary can become a hypertext hub for jumping to parts of interest.

This can be done through passage retrieval [55, 44, 32], probabilistic graphical models, and topic analysis techniques like Latent Dirichlet Allocation (LDA) [4]. To specifically deal with discipline-independent ETDs, topic modeling's generalization capacity can be applied in ETDseer. However, some ETDs are assembled as a collection of scientific papers, which sometimes have relatively loose connections in terms of topics. Such can complicate the use of topic modeling in such ETDs. To overcome this hurdle, we plan to develop new models that will rely on better representations, derived through deep learning.

Research Questions

- How will stakeholders determine which ETD, or chapter of an ETD, or section within a chapter, to read, so that the desired information on a specific topic can be obtained? With ETDs from diverse disciplines, to what extent can we improve the generalization capability of the current techniques in use, and how?

- To summarize key information from different sections of ETDs, how can we improve quality and utility, and from what kinds of aspects? How can we achieve the goal of generating both precise and concise summaries for all hierarchical levels, i.e., in a chapter of an ETD, in an ETD, and in a group or cluster identified from the entire ETD collection?

Research Plan

- *Task 1:* To extract key information from text documents, traditional **probabilistic graphical modeling** techniques have been explored and applied extensively. In these models, words are represented so that each contributes equally in the vocabulary set. These inefficient representations cause generation of imprecise and often highly flawed key information, especially in the context of long ETD documents. We propose to extend the probabilistic graphical modeling approach with distributed representations of words as adopted by word2vec [36] and GloVe [41]. This would produce more efficient representations of words, and result in more meaningful phrases and sentences. Moreover, we will integrate attention models to our modeling mechanisms in order to learn salient words and phrases automatically. Accordingly, we will design a representation learning-based probabilistic graphical model to extract topics and key information from individual sections of the document.

- *Task 2:* To extract a set of **topics from different hierarchical levels** of ETDs, such as in a specific chapter, in a particular ETD, or in an sub-collection of the ETDs, deep learning models such as encoder-decoder or sequence-to-sequence (seq2seq) based approaches will be researched; they have shown promising results in various applications such as in image captioning and machine translation [24]. Furthermore, with additional support of the attention mechanism [38], salient keywords can be learned effectively, which not only ensures high quality of generated multi-topic summarizations but in turn further facilitates our goal in *Task 1* of learning key information. Figure 5 illustrates the conceptual results addressing text summarization problems demonstrated by one of the state of art attention model-based approaches. The general concept is that when generating new words for a summary, at each time step, determine whether to turn to source text or the vocabulary set through a generator gate. If turning to source text for fact generation, the attention model can help decide which words should be paid more attention to at the current time step. We further propose to combine seq2seq and attention mechanisms to train a supervised learning model in ETDseer. Then, we plan to use Amazon Mechanical Turk to evaluate the summaries for ETD documents from different hierarchical levels of content.

2.2.5 Network Visualization

References contained in scholarly publications are used to cite and give credit to related works. The citation links between "citing" and "cited" research works allow easy navigation among related works. Reference networks consisting of citation links not only intuitively present the relationships between research works but also provide a unique way to trace through the evolution of research ideas. Furthermore, different research groups often work on similar research problems. Social networks consisting of author names can suggest potential collaborations between different groups. Readers often want to see such relationships in research communities so that in the future they can pay more attention to these groups. Given the fact that the total number of references in an ETD is significantly larger than in scientific articles, building social and bibliometric networks between papers and papers, papers and ETDs, ETDs and ETDs, as well as their authors, becomes

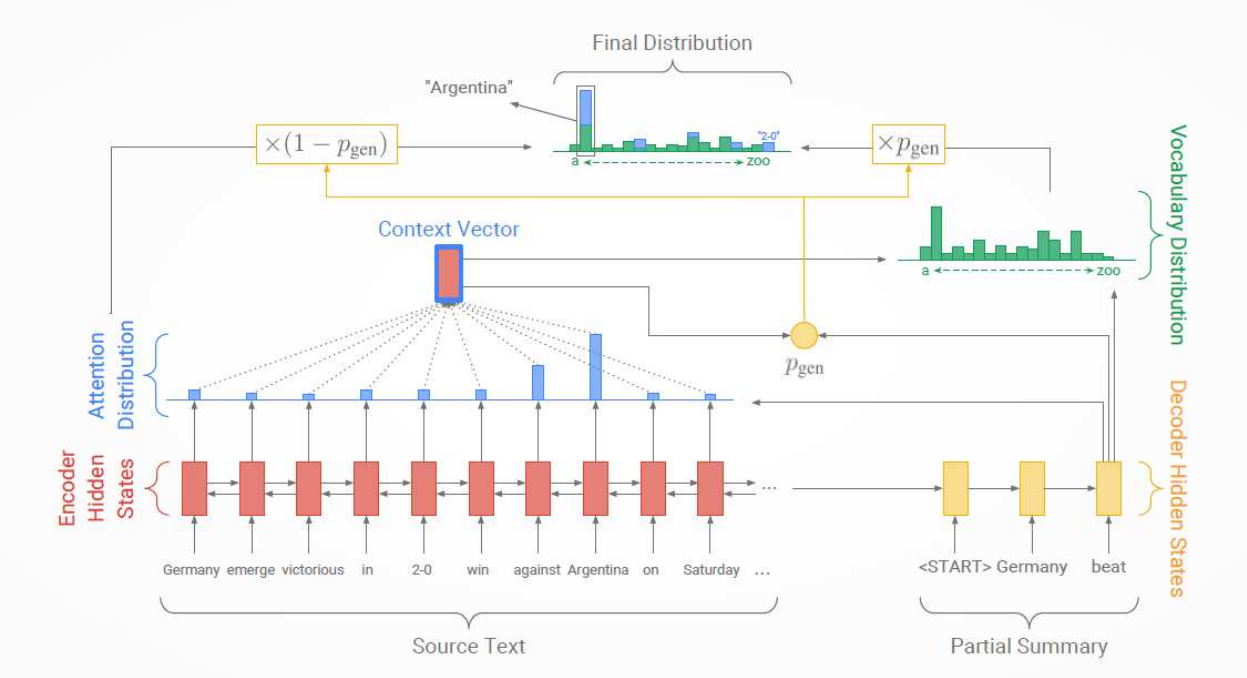

Figure 5: Architecture of state of the art text summarization model [45]

more necessary.

Research Questions

- How can we build a huge network among all the people and references in ETDseer? What are the attributes required for better visualization of the network?
- Considering the resulting social network, how can we characterize the relationships between research groups? What do we need to show in the resulting visualizations?

Research Plan

- *Task 1:* Observing that one reference can cite another but rarely vice versa, we will **build directional reference networks**. We propose to adopt the force-directed graph [18] approach for visualization. In terms of what to show in the network, we will focus on presenting numerical data such as citation counts and paper quality scores.
- *Task 2:* **Social networks** consisting of research groups can reflect their collaboration strengths. These networks can include users like co-authors of the same work, attendees of the same workshop, panelists serving on same panel, and also the research students sharing the same advisor. However, there is no readily populated information indicating the relationships between groups except their mutual citations. To visualize their connections, we need to extract key features such as research similarity. First, we plan to cluster such research groups in terms of their research interests.

After the high-level grouping, fine-grained metrics such as direct citations and research topic similarity based on the particular research problems they work on can be used in visualization.

2.3 Expected Deliverables and their Impacts

We explained the scenarios of how and in what context ETDseer is expected to be used in Section 2.1, and also discussed the system design required to achieve the services described in the scenarios in Section 2.2. In this section we connect the dots and present how our specifically designed system as well as the involved technologies can achieve the desired end results. Figure 6 shows an example of the workflow of the scenario involving a student researcher and the relevant system components we designed for it to be functional. We further list a few other examples of the expected end results of the services we provide by integrating the discussed technologies.

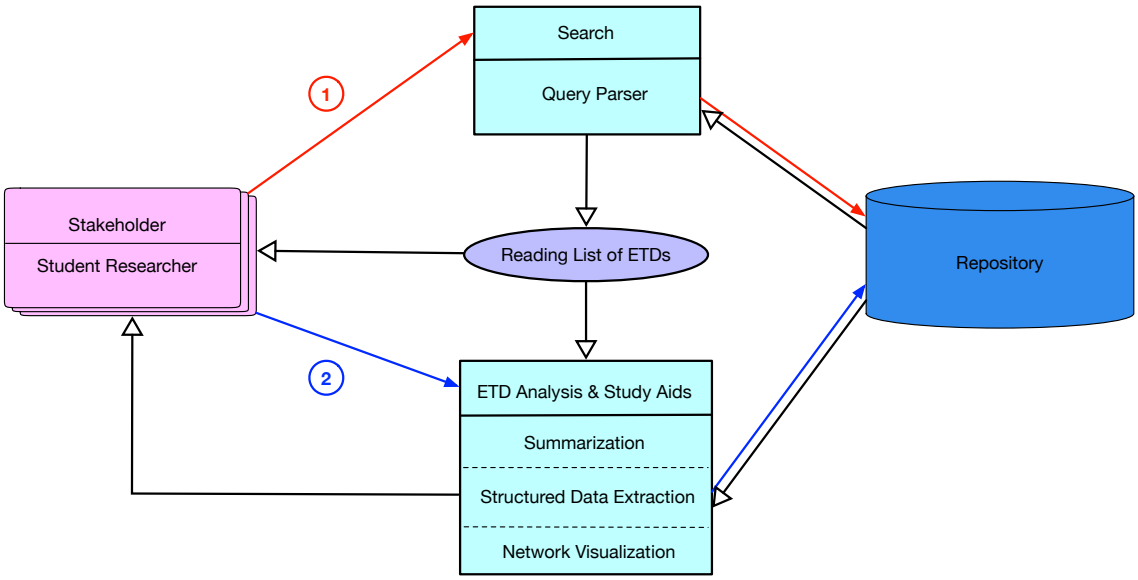

Figure 6: A workflow diagram of a student researcher involved scenario

The following subsections summarize expected deliverables and their impact on some of the key stakeholders.

2.3.1 List of Open Problems and Methods

Using machine learning based text extraction and summarization techniques, ETDseer auto-generates a list of open problems and associated methods for the end users, matching their desired research interests based on the search query. ETDseer provides the functionality of extracting open problems from the *Future Work* and *Conclusion* sections of various ETDs. This service can save user time and effort as critical similarity between their research topic and the methods used in previous works, can lead to waste of time

and resources that the researcher invested. Similarly, ETDseer also extracts methods, approaches, and solutions from *Methodology, Design*, or related sections. This gives the researchers ample options to direct their own approaches in their research. Basically, a customized list of open problems and methods can enable users of ETDseer to explore the novelty of the research problems (through the generated list of open problems), and provide a basis for localizing potential approaches (through the list of the relevant methods) to the problem.

2.3.2 List of ETD metadata

Applying automatic metadata extraction techniques, ETDseer can auto-store a list of metadata fields, and the users can then retrieve the values via searching and browsing. This list consists of features including but not limited to: links to related documents, author information, citation statistics for documents in the provided list, and related reference lists. Related documents would also have the bibliography provided for each corresponding document. A user can utilize it to filter the ETDs that would be most relevant to her interest, while saving significant time and effort to go through all the ETDs without any filter. Extracted author information from the relevant documents can be used to find researchers to collaborate on projects, or contact relevant authors for various purposes like conference invitations for paper submissions and presentations. Citation statistics could be used to measure the quality of the ETDs, and users interested in more popular ETDs can use this feature as well. For researchers and students, finding conferences precisely related to their fields can be a painstaking process, as the list of conferences provided by external websites can be misleading, and not even relevant or of expected quality. ETDseer helps with assessing the quality of the resources it provides, e.g., conference lists, by adding metrics like rank, publication citations, and the percentage of accepted papers.

2.3.3 List of tables, figures, equation, acknowledgement, references

After applying our text extraction and synthesis techniques, a synergistic list of essential parts of ETDs can be generated by ETDseer and presented to the stakeholders in a format conforming to users specific requests. These synergistic lists of information are often very useful to end users. When dealing with hundreds and thousands of files and documents, the summarization part is vital, as it improves efficiency of the system by saving user time. Furthermore, the visualization aspect of summarizing the ETDs is essential, as users can spend minimal time in deciding which of the ETDs and articles in the provided list are worth further exploration. Therefore, ETDseer takes into consideration the visual features of ETDs like tables, figures, and equations, and provides an organized list of summaries accordingly. ETDseer also has the ability to extract information from undervalued sections like acknowledgments, which hold valuable information regarding the organization and people involved in the particular research. Additionally, a comprehensive list of references is provided for the users to explore related works.

3 Related Work

3.1 Academic Citation and Search System

CiteSeerX [6] has been developed at Penn State University by the group led by Dr. Giles, as a public search engine and digital library, and is gaining popularity in scientific and academic paper archiving. Considered to be the first automated citation indexing system (with a patent on this topic), it is the predecessor of academic search tools such as Google Scholar and Microsoft Academic Search, which also focus on short documents. However, it mainly focuses on computer and information science.

3.2 Natural Language Processing

- *Named Entity Recognition* [29, 15]. State-of-the-art NER systems have been developed for well-formed English text. However, due to the varied writing styles among ETDs, it's hard to extend such models to all disciplines. We plan to build an ETD NER dataset and train models for ETDs from different domains to improve NER performance on ETDs.

- *Segmentation*. Most previous segmentation techniques mainly employ heuristic-based strategies [47]. As ETDs are scanned line by line, we can detect the changes of font style or font size, which indicates the start of new sections in a document. Based on such heuristics, hierarchical sections can be segmented. But success depends on how well the ETD is formatted, and ETDs from different disciplines do bear different formatting, which forces development of more generalized segmentation techniques. There are statistics-based models for text segmentation [2], but they rely on the feature representation of the text data. We plan to use recent deep models for feature representations [26] and combine them with general heuristics in all ETDs to boost performance.

- *Topic Analysis*. There have been much work on extracting key texts from paragraphs. Traditional approaches typically use probabilistic graphical models, e.g., extended Latent Dirichlet Allocation [4]. Recent deep learning models, like recurrent neural networks based on attention [38], can help locate salient words/phrases that readers should pay attention to. We will build upon such models with the goal of extracting important texts that can be indexed as metadata to aid user searching.

- *Summarization*. In order to catch the key ideas from long documents, it's essential for the system to summarize them. State-of-the-art methods on text summarization in the NLP field provide good tools that can be used within the framework of the ETD-seer project. Our plan is to rely on recent powerful sequence-to-sequence learning [50] approaches combined with attention mechanisms [38] and memory networks [48] to generate the necessary text summaries. We will build on extensive local work developed by our team, along with other existing methods.

3.3 Image Processing

- *Figure Extraction*. Identifying and extracting figures along with their captions from long documents like ETDs is important both as a way for summarization, and to gain deeper insights into the work. Recent work on Mask R-CNN [26] and its original Region-based CNN [22] for image instance segmentation should be of high relevance to figure extraction. We anticipate building upon these state-of-the-art architectures and incorporating ETD-specific domain knowledge, which will result in a robust figure extraction component.
- *Image Captioning*. Another key aspect for figures in scientific papers is the adjacent captions. Although authors will provide short descriptions about the figures, users want a more elaborate summary about the images. This leads to a requirement for image captioning – an open problem in computer vision. Sequence to sequence learning based models like [68] and [34] are currently exemplifiers of the best approaches. These can help us produce better summaries combined with surrounding texts.

3.4 Table Extraction

Tables are ubiquitous in scientific documents like ETDs. They present experimental results or statistical data in a condensed fashion. However, problems with automatic table extraction from untagged documents, and the lack of a universal table metadata specification, have hindered the success of table processing in digital libraries. TableSeer [33] can crawl digital libraries, detect tables from documents, extract table metadata, and build indexes. We plan to incorporate TableSeer's key components for table manipulation into ETDseer, together with other models [35, 14, 59].

3.5 Visualization

Leveraging the very large social network and reference network produced, visualization is an integral part of the ETDseer system. It can help users gain deep insights about how different groups are working together. Christopher North, a professor at Virginia Tech, is expert on network and digital library visualization [56, 69, 19]. His StarSpire project [5] and V2PI approach for dynamically updating parameters are of high relevance. This will allow more interaction with users as they navigate through the visualized network and learn from it.

4 Expected Significance

Each of the planned series of more functional prototypes of ETDseer will meet the boarder and deeper requirements of various users of the DL system in various scenarios. Documents and references will be analyzed and converted to canonical forms. Figures and tables will be automatically extracted and summarized. The limitations in flexible searching, browsing, visualizing, and exploring of ETDs, and as well as subsections within

ETDs, due to their length and domain complexities, will be significantly minimized, if not eliminated. This research will lead to fundamental contributions in the digital library domain in terms of advanced services discussed in Section 2.1, by developing and applying techniques from information retrieval, information extraction, machine learning, HCI, and AI.

5 Broader Impacts of the Proposed Work

ETDseer, apart from providing the services discussed above, will impact the academic community in a much broader way. It will compute a variety of quality metrics like h-index, citation counts, and conference ranking, to guide users to content that is both relevant and of high value. Negative results, that rarely are published, will become available for the first time. Details of research – given in ETD text, figures, and tables that rarely appear elsewhere – will become accessible, and ensure a deeper understanding of methods and findings. While this advanced visualization of data can enhance the overall user experience, it can also be highly beneficial for users who have difficulty reading and interpreting works in general. A rich range of services connected with hundreds of thousands of (multimedia) ETDs, will dramatically advance research by graduate students (who make frequent use of others' theses) and the broader scholarly community, which requires an efficient system that enables more research and helps produce important results in less time.

References

[1] The NDLTD Union Catalog. `http://union.ndltd.org/portal/`. OCLC Research.

[2] Doug Beeferman, Adam Berger, and John Lafferty. Statistical models for text segmentation. *Machine learning*, 34(1):177–210, 1999.

[3] Sumit Bhatia, Cornelia Caragea, Hung-Hsuan Chen, Jian Wu, Pucktada Treeratpituk, Zhaohui Wu, Madian Khabsa, Prasenjit Mitra, and C. Lee Giles. Specialized research datasets in the CiteSeerX digital library. *D-Lib Magazine*, 18(7/8), 2012.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning research*, 3(Jan):993–1022, 2003.

[5] Lauren Bradel, Chris North, and Leanna House. Multi-model semantic interaction for text analytics. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pages 163–172. IEEE, 2014.

[6] Cornelia Caragea, Jian Wu, Alina Ciobanu, Kyle Williams, Juan Fernández-Ramírez, Hung-Hsuan Chen, Zhaohui Wu, and Lee Giles. CiteseerX: A scholarly big dataset. In *European Conference on Information Retrieval*, pages 311–322. Springer, 2014.

[7] Hung-Hsuan Chen, Liang Gou, Xiaolong Zhang, and Clyde Lee Giles. Collabseer: a search engine for collaboration discovery. In *Proceedings of the 2011 Joint International Conference on Digital Libraries, JCDL 2011, Ottawa, ON, Canada, June 13-17, 2011*, pages 231–240, 2011.

[8] Hung-Hsuan Chen, Pucktada Treeratpituk, Prasenjit Mitra, and C. Lee Giles. CSSeer: an expert recommendation system based on CiteSeerX. In *13th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '13, Indianapolis, IN, USA, July 22 - 26, 2013*, pages 381–382, 2013.

[9] Jianpeng Cheng and Dimitri Kartsaklis. Syntax-aware multi-sense word embeddings for deep compositional models of meaning. *EMNLP*, 2015.

[10] Sagnik Ray Choudhury and Clyde Lee Giles. An architecture for information extraction from figures in digital libraries. In *WWW (Companion Volume)*, pages 667–672, 2015.

[11] Sagnik Ray Choudhury, Prasenjit Mitra, Andi Kirk, Silvia Szep, Donald Pellegrino, Sue Jones, and C Lee Giles. Figure metadata extraction from digital documents. In *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*, pages 135–139. IEEE, 2013.

[12] Sagnik Ray Choudhury, Shuting Wang, and C Lee Giles. Curve separation for line graphs in scholarly documents. In *Digital Libraries (JCDL), 2016 IEEE/ACM Joint Conference on*, pages 277–278. IEEE, 2016.

[13] Sagnik Ray Choudhury, Shuting Wang, and C Lee Giles. Scalable algorithms for scholarly figure mining and semantics. In *SBD@ SIGMOD*, page 1, 2016.

[14] Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74, 2008.

[15] Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S Weld, and Alexander Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artificial intelligence*, 165(1):91–134, 2005.

[16] Umer Farooq, Craig H. Ganoe, John M. Carroll, and C. Lee Giles. Designing for e-science: Requirements gathering for collaboration in CiteSeer. *Int. J. Hum.-Comput. Stud.*, 67(4):297–312, 2009.

[17] E. Fox, R. Hall, and N. Kipp. NDLTD: preparing the next generation of scholars for the information age. *The New Review of Information Networking (NRIN)*, pages 59–76, 1997.

[18] Thomas MJ Fruchterman and Edward M Reingold. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164, 1991.

[19] Emden R Gansner and Stephen C North. An open graph visualization system and its applications to software engineering. *Software Practice and Experience*, 30(11):1203–1233, 2000.

[20] C. Lee Giles. The future of CiteSeer: CiteSeerX. In *Knowledge Discovery in Databases: PKDD 2006, 10th European Conference on Principles and Practice of Knowledge Discovery in Databases, Berlin, Germany, September 18-22, 2006, Proceedings*, page 2, 2006.

[21] C Lee Giles, Kurt D Bollacker, and Steve Lawrence. CiteSeer: An automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries*, pages 89–98. ACM, 1998.

[22] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[23] M.A. Gonalves, E.A. Fox, L.T. Watson, and N. Kipp. Streams, structures, spaces, scenarios, societies (5S): a formal model for digital libraries. In *ACM Transactions on Information Systems, 22 (2)*, pages 270–312, 2004.

[24] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.

[25] Hui Han, C. Lee Giles, Eren Manavoglu, and Hongyuan Zha. Automatic document metadata extraction using Support Vector Machines. In *Proceedings of the 2003 Joint Conference on Digital Libraries (JCDL03)*, 2003.

[26] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. *arXiv preprint arXiv:1703.06870*, 2017.

[27] Wenyi Huang, Zhaohui Wu, Prasenjit Mitra, and C. Lee Giles. RefSeer: A citation recommendation system. In *IEEE/ACM Joint Conference on Digital Libraries, JCDL 2014, London, United Kingdom, September 8-12, 2014*, pages 371–374, 2014.

[28] Madian Khabsa, Pucktada Treeratpituk, and C. Lee Giles. AckSeer: a repository and search engine for automatically extracted acknowledgments from digital libraries. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '12, Washington, DC, USA, June 10-14, 2012*, pages 185–194, 2012.

[29] Wai Lam, Shing-Kit Chan, and Ruizhang Huang. Named entity translation matching and learning: With application for mining unseen translations. *ACM Transactions on Information Systems (TOIS)*, 25(1):2, 2007.

[30] Huajing Li, Isaac G. Councill, Levent Bolelli, Ding Zhou, Yang Song, Wang-Chien Lee, Anand Sivasubramaniam, and C. Lee Giles. CiteSeerX: a scalable autonomous scientific digital library. In *Proceedings of the 1st international conference on Scalable information systems*, InfoScale '06, New York, NY, USA, 2006. ACM.

[31] Huajing Li, Isaac G. Councill, Wang-Chien Lee, and C. Lee Giles. CiteSeerX: an architecture and web service design for an academic document search engine. In *Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23-26, 2006*, pages 883–884, 2006.

[32] Xiaoyong Liu and W Bruce Croft. Passage retrieval based on language models. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 375–382. ACM, 2002.

[33] Ying Liu, Kun Bai, Prasenjit Mitra, and C Lee Giles. TableSeer: automatic table metadata extraction and searching in digital libraries. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries (JCDL2007)*, pages 91–100, Vancouver, British Columbia, Canada, June 17–22, 2007.

[34] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *arXiv preprint arXiv:1612.01887*, 2016.

[35] Simone Marinai. Metadata extraction from PDF papers for digital library ingest. In *Document Analysis and Recognition, 2009. ICDAR'09. 10th International Conference on*, pages 251–255. IEEE, 2009.

[36] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[37] Elke Mittendorf and Peter Schäuble. Document and passage retrieval based on Hidden Markov Models. In *SIGIR94*, pages 318–327. Springer, 1994.

[38] Chris Olah and Shan Carter. Attention and augmented recurrent neural networks. *Distill*, 1(9):e1, 2016.

[39] Alexander G. Ororbia, Jian Wu, Madian Khabsa, Kyle Williams, and Clyde Lee Giles. Big scholarly data in CiteSeerX: Information extraction from the web. In *Proceedings of the 24th International Conference on World Wide Web Companion, WWW 2015, Florence, Italy, May 18-22, 2015 - Companion Volume*, pages 597–602, 2015.

[40] S. Park and E. Fox. Enriching the VT ETD-db system with references. In *Proceeding of 14th International Symposium on Electronic Theses and Dissertations. NDLTD*, 2011.

[41] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[42] David Pinto, Andrew McCallum, Xing Wei, and W. Bruce Croft. Table extraction using Conditional Random Fields. In *SIGIR '03 Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, Toronto, Canada July 28 - August 01*, pages 235–242, 2003.

[43] Sagnik Ray Choudhury, Prasenjit Mitra, and Clyde Lee Giles. Automatic extraction of figures from scholarly documents. In *Proceedings of the 2015 ACM Symposium on Document Engineering*, pages 47–50. ACM, 2015.

[44] Gerard Salton, James Allan, and Chris Buckley. Approaches to passage retrieval in full text information systems. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 49–58. ACM, 1993.

[45] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

[46] Richard Socher, Eric H Huang, Jeffrey Pennington, Andrew Y Ng, and Christopher D Manning. Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *NIPS*, volume 24, pages 801–809, 2011.

[47] V. Srinivasan, M. Magdy, and E. Fox. Enhanced browsing system for Electronic Theses and Dissertations. In *Proceeding of 14th International Symposium on Electronic Theses and Dissertations. NDLTD*, 2011.

[48] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448, 2015.

[49] Bingjun Sun, Prasenjit Mitra, and C. Lee Giles. Mining, indexing, and searching for textual chemical molecule information on the web. In *WWW 2008 / Refereed Track: Web Engineering - Applications, April 21-25, 2008, Beijing, China*, 2008.

[50] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[51] Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 41–47. ACM, 2003.

[52] Pradeep Teregowda and C. Lee Giles. Scaling SeerSuite in the Cloud. In *Proceedings of the 2013 IEEE International Conference on Cloud Engineering*, IC2E '13, pages 146–155, Washington, DC, USA, 2013. IEEE Computer Society.

[53] Pradeep B. Teregowda, Isaac G. Councill, R. Juan Pablo Fernández, Madian Khabsa, Shuyi Zheng, and C. Lee Giles. SeerSuite: developing a scalable and reliable application framework for building digital libraries by crawling the web. WebApps'10, pages 14–14, 2010.

[54] Pradeep B. Teregowda, Bhuvan Urgaonkar, and C. Lee Giles. CiteSeerX: a cloud perspective. In *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, HotCloud'10, pages 9–9, Berkeley, CA, USA, 2010.

[55] Courtney Wade and James Allan. Passage retrieval and evaluation. Technical report, DTIC Document, 2005.

[56] Jun Wang, Abhishek Agrawal, Anil Bazaza, Supriya Angle, Edward A Fox, and Chris North. Enhancing the Envision interface for digital libraries. In *Proceedings of the 2nd ACM/IEEE-CS joint conference on Digital libraries*, pages 275–276. ACM, 2002.

[57] K. Williams, J. Wu, S. R. Choudhury, M. Khabsa, and C. L. Giles. Scholarly big data information extraction and integration in the CiteSeerX digital library. In *Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference on*, pages 68–73, March 2014.

[58] Kyle Williams, Jian Wu, and C. Lee Giles. SimSeerX: a similar document search engine. In *ACM Symposium on Document Engineering 2014, DocEng '14, Fort Collins, CO, USA, September 16-19, 2014*, pages 143–146, 2014.

[59] Ian H Witten, David Bainbridge, Gordon Paynter, and Stefan Boddie. Importing documents and metadata into digital libraries: Requirements analysis and an extensible architecture. In *International Conference on Theory and Practice of Digital Libraries*, pages 390–405. Springer, 2002.

[60] J. Wu, P. Teregowda, K. Williams, M. Khabsa, D. Jordan, E. Treece, Z. Wu, and C. L. Giles. Migrating a digital library to a private cloud. In *Cloud Engineering (IC2E), 2014 IEEE International Conference on*, pages 97–106, March 2014.

[61] Jian Wu, Chen Liang, Huaiyu Yang, and C. Lee Giles. CiteSeerX data: Semanticizing scholarly papers. In *Proceedings of the International Workshop on Semantic Big Data*, SBD '16, pages 2:1–2:6, New York, NY, USA, 2016. ACM.

[62] Jian Wu, Alexander Ororbia, Kyle Williams, Madian Khabsa, Zhaohui Wu, and C. Lee Giles. Utility-based control feedback in a digital library search engine: Cases in CiteSeerX. In *9th International Workshop on Feedback Computing, Philadelphia, PA, USA, June 17, 2014.*, 2014.

[63] Jian Wu, Pradeep Teregowda, Madian Khabsa, Stephen Carman, Douglas Jordan, Jose San Pedro Wandelmer, Xin Lu, Prasenjit Mitra, and C. Lee Giles. Web crawler middleware for search engine digital libraries: a case study for CiteSeerX. WIDM '12, pages 57–64, New York, NY, USA, 2012. ACM.

[64] Jian Wu, Pradeep Teregowda, Madian Khabsa, Eric Treece, Douglas Jordan, Stephen Carman, Prasenjit Mitra, and C. Lee Giles. Scalability bottlenecks of the CiteSeerX digital library search engine. In *Proceedings of Large-Scale And Distributed Systems for Information Retrieval (WSDM'12)*, October 2012.

[65] Jian Wu, K. Williams, M. Khabsa, and C.L. Giles. The impact of user corrections on a crawl-based digital library: A CiteSeerX perspective. In *Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom), 2014 International Conference on*, pages 171–176, Oct 2014.

[66] Jian Wu, Kyle Williams, Hung-Hsuan Chen, Madian Khabsa, Cornelia Caragea, Alexander Ororbia, Douglas Jordan, and C. Lee Giles. CiteSeerX: AI in a digital library search engine. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, pages 2930–2937, 2014.

[67] Jian Wu, Kyle Mark Williams, Hung-Hsuan Chen, Madian Khabsa, Cornelia Caragea, Suppawong Tuarob, Alexander G. Ororbia, Douglas Jordan, Prasenjit Mitra, and C. Lee Giles. CiteSeerX: AI in a digital library search engine. *AI Magazine*, 36:35–48, 2015.

[68] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81, 2015.

[69] Beth Yost and Chris North. The perceptual scalability of visualization. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):837–844, 2006.