Scenarios for Advanced Services in an ETD Digital Library

Yufeng Ma[1], Tingting Jiang[1], Chandani Shrestha[1], Edward A. Fox[1], Jian Wu[2], C. Lee Giles[2]

Virginia Tech[1]

Pennsylvania State University[2]

{yufengma, virjtt03, chandani, fox}@vt.edu, {jxw394, giles}@ist.psu.edu

Abstract

In this paper, we describe scenarios involving various groups of stakeholders who can benefit from using a digital library specialized to work with electronic theses and dissertations (ETDs). Readers also should gain insight into what advanced services might aid those interested in ETDs to engage in such scenarios, and what machine learning approaches could support those services.

Basic digital libraries support indexing, searching, and browsing. Even those services, however, are limited when only metadata is available. Using full text to extend faceted searching provides improvement, but adds noise and reduces precision. Natural language processing, applied to information extraction, yields additional improvement, whereby authors, dates, locations, organizations, and other entities are identified and added to metadata records and facets, supporting specificity and linking. Digital libraries like CiteSeerX, and services like Google Scholar -- which extract, analyze, and link references in publications -- provide additional capabilities that would be valuable for ETDs, but there has not been dedicated work to verify if existing methods are applicable to ETDs (due to length, complexity, and domain variations).

ETDs, typically in PDF, are a largely untapped international resource. Digital libraries with tailored services could effectively address the broad needs to discover and utilize available ETDs of interest. We are researching a tailored digital library for English ETDs, which would offer special services for large ETD collections: review the literature, identify hypotheses, list research questions, explain approaches, describe methods, summarize results, discuss findings, present conclusions, and provide insights about open problems. We could 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 could be prototyped now, using advanced natural language processing and information extraction techniques, coupled with machine learning and information retrieval methods. The resulting system, called ETDseer, would enable stakeholders to engage in advanced scenarios -- like those discussed in this paper -- that go well beyond conventional searching and browsing.

Keywords: CiteSeerX, deep learning, ETDseer, information extraction (IE), information retrieval (IR), machine learning (ML), natural language processing (NLP), NDLTD, scenarios

1. Introduction

ETDs can be a valuable aid to learning and scholarship. Thanks to the efforts of many students, faculty, administrators, colleges, universities, libraries, consortia, supporting organizations, and technologists, there now are some 5 million ETDs in the Networked Digital Library of Theses and Dissertations (NDLTD) Union Catalog. Figure 1 illustrates how basic faceted browsing and searching of those works is supported through the NDLTD Global Search service.

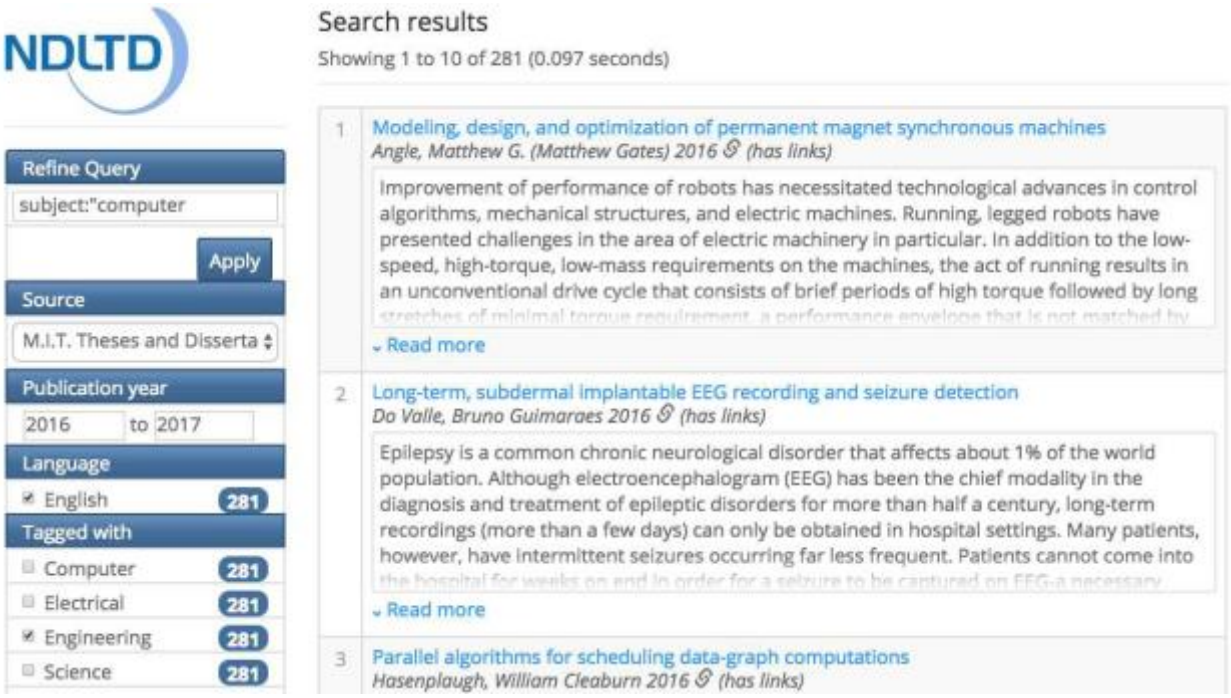

Figure 1. NDLTD Global Search example

While it is important to improve further regarding the number, expressive nature, diversity of content (e.g., including multimedia formats, datasets, software, hypermedia links, etc.), and accessibility of ETDs, the existing collections already constitute a largely unused but highly valuable international resource. Tapping this resource requires research both on more effective ways to discover ETDs of interest, and on 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. One way forward is illustrated by systems like Google Scholar and CiteSeerX (Caragea et al. 2014, Giles 2006, Giles, Bollacker, and Lawrence 1998, Li et al. 2006, Ororbia II et al. 2015, Teregowda, Urgaonkar, and Giles 2010), that extract, analyze, and link references in academic documents, mostly conference and journal articles; see Figure 2. Documents, their metadata, and citations, are available in search results of Google Scholar and CiteSeerX. Some digital libraries such as Semantic Scholar (https://www.semanticscholar.org/) also show figures and tables. However, such currently available systems do not work well with ETDs (due to length, complexity, and domain variations).

Though there are widely available Digital Libraries (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 the most efficient way to search large files in natural language. One promising way is to build a topical model by extracting key phrases using a combination of natural language processing (NLP) and artificial intelligence (AI). Accordingly, in this paper, we discuss the need for a tailored DL focusing on ETDs based on advanced concepts from NLP, information extraction (IE), information retrieval (IR), Web crawling/archiving, and current breakthroughs from machine learning (ML) and deep learning. Hence, the next sections identify requirements for such systems, described through scenarios, and more about approaches that might aid in building these systems.

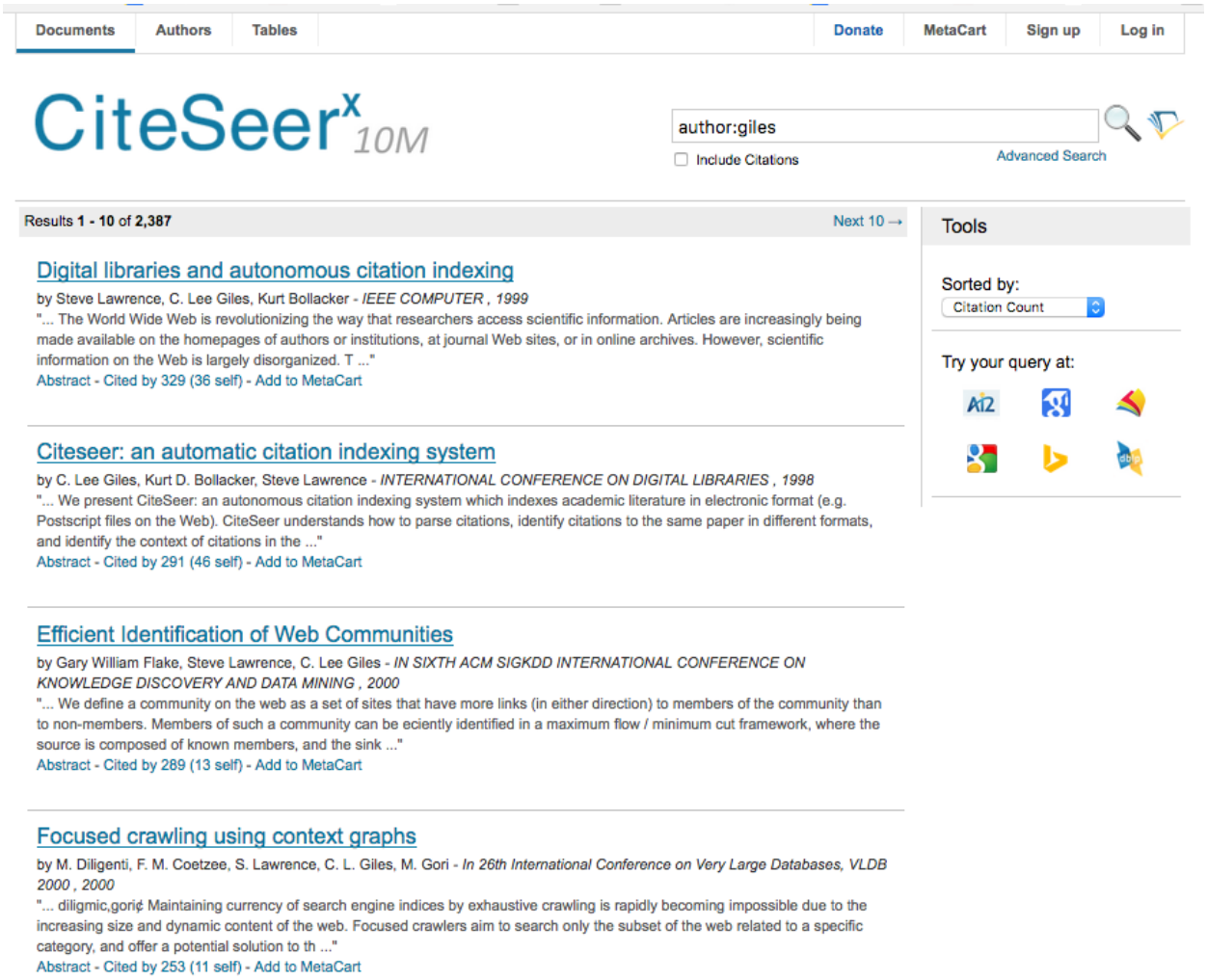

Figure 2. CiteSeerX example

2. Advanced Scenarios

Based on previous digital library developments and current technologies, we are 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. According to 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.

Table 1: Advanced Scenarios Utilizing ETDs

Stakeholders	**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 Researcher	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
	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/bibliometric 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 approached 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 (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
Journal Editor	Peer-reviewer ID	Research interest-based reviewer list
	Content originality check	Previous publications of the authors Estimated percentage of new content/work

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). Clusters show 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 approaches have been employed, but without details and comparative studies. An advanced ETD DL can help SR identify the ETDs that are related to each of the methods as well as corresponding involved datasets. Then it lists pointers to descriptions of the source code, as well as to 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, a detailed outline of a paper derived from her 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. The ETD system provides a report 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 from the ETD DL. Based on the ETDs of interest, it 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 the system 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, it 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 the ETD DL a list of topics from the announcement. It 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, the ETD system 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 our system with a list of keywords, related to the theme of the conference. It 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 can query an ETD DL using keywords from the submission. Then it 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 our system. JE also needs to check if the paper submission has at least 30% original content relative to previous publications. JE also queries our system with the author names for an originality check. It then 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.

3. Key Approaches

The following subsections sketch four broad approaches to building services in a DL that could support the advanced scenarios described in Section 2.

Building upon Existing DL Technologies

A proper integration of the DLs discussed above in Section 1 (see Figures 1, 2), where the datasets (metadata and long documents) are mostly from NDLTD, and technical features come from CiteSeerX, could guide the initial development of a system called ETDseer. ETDseer would leverage most of the DL tools that CiteSeerX has provided. Although CiteSeerX can work as a technological foundation for building ETDseer, the target documents that CiteSeerX serves are mostly conference and journal articles. 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 are not sufficient. Furthermore, unlike CiteSeerX, most of the functional requirements in ETDseer, such as extracting research questions and hypotheses, cannot be addressed via keyword and full text only searching and browsing.

CiteSeerX handles documents from several specific scientific disciplines, mostly from computer science, mathematics, physics, chemistry, and recently from medical science. The writing format and styles of these domain specific documents are fairly consistent, so a database can be built using a set of universal extraction and parsing tools. However, ETDs can be from a more diverse range of 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 deep learning methods, can potentially be used to achieve the advanced functionalities of ETDseer.

Structured Data Extraction

Since ETDs are long documents, it is appropriate to enhance document retrieval with passage retrieval. Furthermore, advanced ETD DLs 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 typical scientific papers, in which the authors usually arrange contents by 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, while others are presented as a composite of scientific papers where each chapter represents a complete work. 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. Existing CiteSeerX-type segmentation and extraction tools are not flexible when working with such varying formats, or with unpredictable structures.

However, segmentation of ETD documents might be achieved through a combination of heuristic-based strategies (Srinivasan, Magdy, and Fox 2011) and deep learning approaches. Heuristic methods could leverage positional, layout, font, style, and numbering information. To aid segmentation, deep learning approaches might treat documents as pictures, so segmentation or detection approaches like Mask R-CNN (He et al. 2017) could be applied.

Tables and figures are effective devices for presenting research results. It is especially important to extract these types of structured data when dealing with scientific data in fields like computer science and chemistry. Accordingly, extensions to the technologies used in CiteSeerX have led to a system, TableSeer (Liu et al. 2007), and research studies of figures (Ray Choudhury and Giles 2015, Choudhury et al. 2013, Choudhury, Wang, and Giles 2016a, b, Ray Choudhury, Mitra, and Giles 2015, Williams et al. 2014). With discipline independent ETDs, the content in figures and tables, likely to be domain specific, would be more difficult to extract. Accordingly, enhanced machine-learning based approaches should be applied in addition to existing CiteSeerX methods, e.g., Clark et al. 2016.

Existing citation extraction tools, e.g., ParsCit (Councill, Giles, and Kan 2008), from CiteSeerX, perform well on extracting references at the end of documents and thus they can be directly applied to some ETDs. For references occurring at the end of each chapter, or in other locations like footnotes or table entries, new machine learning-based approaches would be required to locate the reference sections before they are parsed. These would require new feature extraction as well as fine-grained feature selection, to improve accuracy and efficiency. Classifiers could be trained to identify references appearing at unusual locations. Extracted references ultimately would be converted into some canonical format like BibTeX so they could be presented in any user-desired style.

Text Generation

Other types of presentations might include summaries. A good summary could become a hypertext hub for jumping to parts of interest. These could be generated through passage retrieval (Liu and Croft 2002, Salton, Allan, and Buckley 1993, Wade and Allan 2005), extensions to probabilistic graphical models via word2vec (Mikolov et al. 2013) or GloVe (Pennington, Socher, and Manning 2014), and topic analysis techniques like Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003). To specifically deal with discipline-independent ETDs, topic modeling's generalization capacity should be employed. However, some ETDs are assembled as a collection of scientific papers, which sometimes have relatively loose connections in terms of topics. Such ETDs can complicate the use of topic modeling. To overcome this hurdle, we plan to develop new models that will rely on better representations, derived through deep learning.

To extract a set of topics from different hierarchical levels of ETDs, such as in a specific chapter, in a particular ETD, or in a sub-collection of the ETDs, deep learning models such as encoder-decoder or sequence-to-sequence based approaches (Sutskever, Vinyals, and Le 2014) might be adapted; they have shown promising results in image captioning and machine translation (Goodfellow, Bengio, and Courville 2016). Furthermore, with additional support of the attention mechanism (Olah and Carter 2016), models could be trained to extract key phrases, which could facilitate multi-topic summarization.

Network Visualization

References contained in scholarly publications are used to cite and give credit to related works. 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 more necessary. Hence we would build directional reference networks and employ a force-directed (Fruchterman and Reingold 1991) graph approach for visualization. In terms of what to show in the network, we could focus on presenting numerical data such as citation counts (reflecting collaboration potential) and paper quality scores.

These networks can include users like co-authors of the same work, attendees of the same workshop, speakers serving on the same panel, and research students sharing the same advisor. However, since there is little information indicating the relationships between groups except their mutual citations, we could estimate research similarity based on clustering in terms of research interests. After a high-level grouping, fine-grained metrics such as direct citations and research topic similarity based on the particular research problems they work on could be used in visualization.

4. Conclusion

A tailored DL could help the many different stakeholders who could benefit from the current and emerging large collections of online open access ETDs. Section 1 introduces key issues and explains the current situation regarding ETD collections and systems. Section 2 gives an overview of scenarios wherein a wide range of users could benefit from advanced DL services for ETDs. Section 3 describes four aspects of a possible DL, ETDseer, giving examples of promising techniques and approaches that could lead to a novel tailored DL for ETDs. It is hoped that research can build upon this discussion, so that the vast potential hidden in the growing global collection of ETDs can help with research, education, and scholarly activities.

5. Acknowledgements

We gratefully acknowledge partial support from the NSF through grant IIS-1423337, and from IMLS through grant LG-71-16-0037-16.

6. References

Allen, Institute. 2017. "Semantic Scholar." accessed 07/31/2017. https://www.semanticscholar.org/.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. "Latent dirichlet allocation." *Journal of machine Learning research* 3 (Jan):993-1022.

Caragea, Cornelia, Jian Wu, Alina Maria Ciobanu, Kyle Williams, Juan Pablo Fernández Ramírez, Hung-Hsuan Chen, Zhaohui Wu, and C Lee Giles. 2014. "CiteSeerX: A Scholarly Big Dataset." ECIR.

Choudhury, Sagnik Ray, Prasenjit Mitra, Andi Kirk, Silvia Szep, Donald Pellegrino, Sue Jones, and C Lee Giles. 2013. "Figure metadata extraction from digital documents." Document Analysis and Recognition (ICDAR), 2013, 12th International Conference on.

Choudhury, Sagnik Ray, Shuting Wang, and C Lee Giles. 2016a. "Curve separation for line graphs in scholarly documents." Digital Libraries (JCDL), 2016, IEEE/ACM Joint Conference on.

Choudhury, Sagnik Ray, Shuting Wang, and C Lee Giles. 2016b. "Scalable algorithms for

scholarly figure mining and semantics." SBD@ SIGMOD.

Clark, Christopher, and Santosh Divvala. 2016. "Pdffigures 2.0: Mining figures from research papers." Digital Libraries (JCDL), 2016, IEEE/ACM Joint Conference on.

Councill, Isaac G, C Lee Giles, and Min-Yen Kan. 2008. "ParsCit: an Open-source CRF Reference String Parsing Package." LREC.

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

Giles, C Lee. 2006. "The future of CiteSeer: CiteSeerX." Proceedings of the 10th European conference on Principle and Practice of Knowledge Discovery in Databases.

Giles, C Lee, Kurt D Bollacker, and Steve Lawrence. 1998. "CiteSeer: An automatic citation indexing system." Proceedings of the third ACM conference on Digital libraries.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. 2016. *Deep learning*: MIT Press.

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

Li, Huajing, Isaac G Councill, Levent Bolelli, Ding Zhou, Yang Song, Wang-Chien Lee, Anand Sivasubramaniam, and C Lee Giles. 2006. "CiteSeerX: a scalable autonomous scientific digital library." Proceedings of the 1st international conference on Scalable information systems.

Liu, Xiaoyong, and W Bruce Croft. 2002. "Passage retrieval based on language models." Proceedings of the eleventh international conference on Information and knowledge management.

Liu, Ying, Kun Bai, Prasenjit Mitra, and C Lee Giles. 2007. "TableSeer: automatic table metadata extraction and searching in digital libraries." Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems.

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

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." Proceedings of the 24th International Conference on World Wide Web.

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

Ray Choudhury, Sagnik, and Clyde Lee Giles. 2015. "An architecture for information extraction from figures in digital libraries." Proceedings of the 24th International Conference on World Wide Web.

Ray Choudhury, Sagnik, Prasenjit Mitra, and Clyde Lee Giles. 2015. "Automatic extraction of figures from scholarly documents." Proceedings of the 2015 ACM Symposium on Document Engineering.

Salton, Gerard, James Allan, and Chris Buckley. 1993. "Approaches to passage retrieval in full text information systems." Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval.

Srinivasan, Venkat, Mohamed Magdy, and Edward A Fox. 2011. "Enhanced browsing system for Electronic Theses and Dissertations." Proceeding of 14th International Symposium on Electronic Theses and Dissertations. NDLTD.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. 2014. "Sequence to sequence learning with neural networks." Advances in neural information processing systems.

Teregowda, Pradeep B, Bhuvan Urgaonkar, and C Lee Giles. 2010. "CiteSeerX: A Cloud Perspective." HotCloud.

Wade, Courtney, and James Allan. 2005. Passage retrieval and evaluation. University of Massachusetts, Amherst, Center for Intelligent Information Retrieval.

Williams, Kyle, Jian Wu, Sagnik Ray Choudhury, Madian Khabsa, and C Lee Giles. 2014. "Scholarly big data information extraction and integration in the CiteseerX digital library." Data Engineering Workshops (ICDEW), 2014, IEEE 30th International Conference on.

Author Biographies

Yufeng Ma received the BE degree in Computer Science from Wuhan University, China in 2012. Currently he is a Ph.D. student in the Department of Computer Science at Virginia Tech, USA, where he is working on connecting natural language with images with deep learning approaches. His research interests include text data mining, information retrieval, computer vision, deep learning, and artificial intelligence in general.

Tingting Jiang is working on her Ph.D. degree in Computer Science at Virginia Tech while working full-time as a Software Engineer at Virginia Tech University Libraries. Her research area is in wireless networking and cyber security. She was a recipient of an NSF Graduate Research Fellowship and a Microsoft Research Graduate Women's Scholarship. During 2007-2009, she was a Software Engineer at Intrexon Corporation, Blacksburg, VA.

Chandani Shrestha is a Ph.D. candidate at Virginia Tech, Computer Science department, where she holds a position of Teaching Assistant. Her primary research interests include HCI, which she is pursuing at the THIRD Lab at Virginia Tech. She completed her undergraduate degree, majoring in Computer Science with honors, from Benedict College, SC. She was a recipient of a Trustee as well as International Service Scholar scholarship at Benedict College. She founded a CS basic learning project for young girls in Nepal, called KM Computing for Girls, for which she was awarded the Systers Pass-It-On award.

Edward A. Fox is a Professor of Computer Science (with courtesy appointment in ECE) at Virginia Tech, where he also is Director of the Digital Library Research Laboratory (DLRL). For the Networked Digital Library of Theses and Dissertations he serves as Executive Director and Chairman of the Board. He is a Fellow of the IEEE, and served on the Board of Directors of the Computing Research Association. He first became interested in ETDs in 1987, and continues to explore technical, policy, and administrative matters related. Multiple ETD-related student and sponsored research projects have involved the DLRL, which engages in work with information, including digital libraries, information retrieval, machine learning, and human-computer interaction.

Jian Wu is a Lecturer at the College of Information Sciences and Technology at the Pennsylvania State University. He received his Ph.D. from the Department of Astronomy and Astrophysics at the Pennsylvania State University in 2011. Dr. Wu has been pursuing research on data science and technologies focusing on applied machine learning on mining textual information in

scientific documents, including but not limited to classification, extraction, and search. He published more than 20 peer-reviewed papers in IEEE, ACM, and AAAI conferences with one best paper and two best paper nominations. He also published articles in top astrophysical journals. He is the technical leader of the world-renown CiteSeerX project since 2013, and now is co-leading this project with Dr. C. Lee Giles. Dr. Wu teaches undergraduate level database and programming language courses and co-teaches the Information Retrieval class.

Dr. C. Lee Giles is the David Reese Professor at the College of Information Sciences and Technology at the Pennsylvania State University, University Park, PA, with a graduate school appointment in Computer Science and Engineering. He is a Fellow of the ACM, IEEE, and INNS (Gabor prize). He is known for the digital library search engine, CiteSeer, which he co-created, developed, and maintains plus other related search engines. He has expertise in information and metadata extraction and indexing, data extraction, name matching and linking, all at scale. He has published over 400 papers in these areas with over 30,000 citations and an h-index over 80.