ETDseer

Final Term Term Project Presentation

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

Overview

- **Problem Statements/Motivation**
- **Related Works**
  - NDLTD
  - CiteSeer
- **Stakeholder Overview**

  - **Architecture**
  - **Scenarios and Services**
    - Workflow example

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

      - **Working On…**

Background

Problem Statements

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

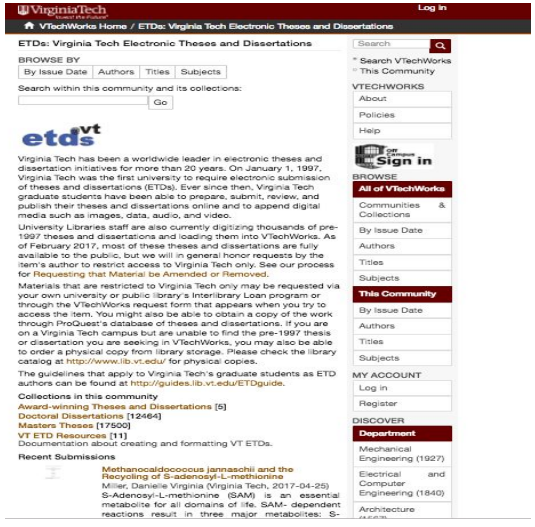
Motivation

- Rich knowledge base - Single platform
- Accessible to broader group of users
- Network of institutional repositories

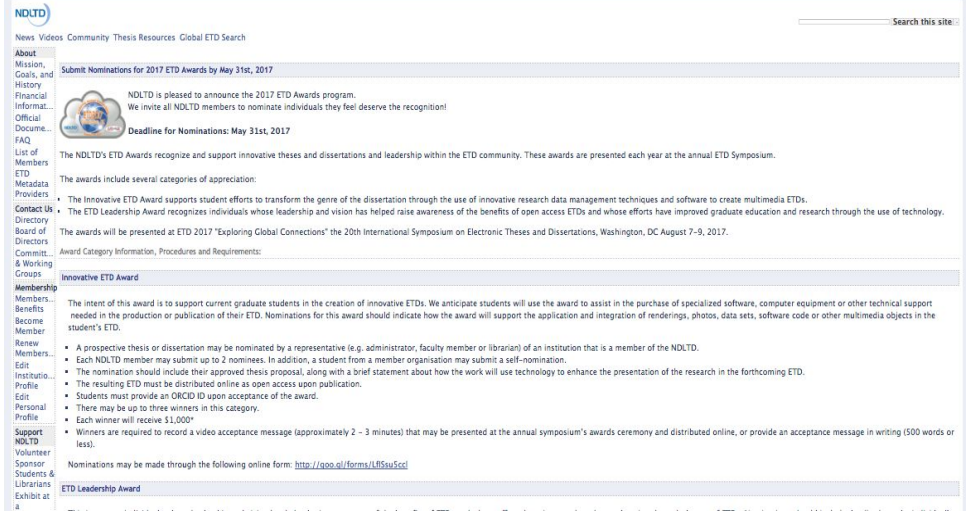
Related Works

- NDLTD
- VTechworks
- CiteSeer
- ContentMine

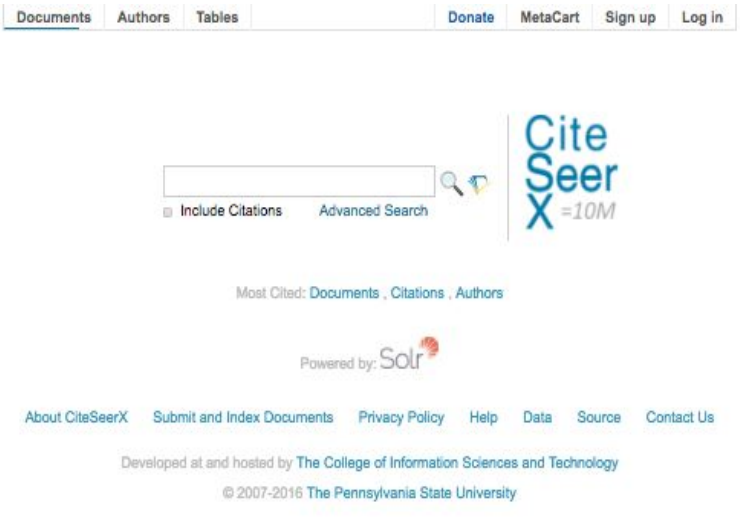
VTechWorks



NDLTD

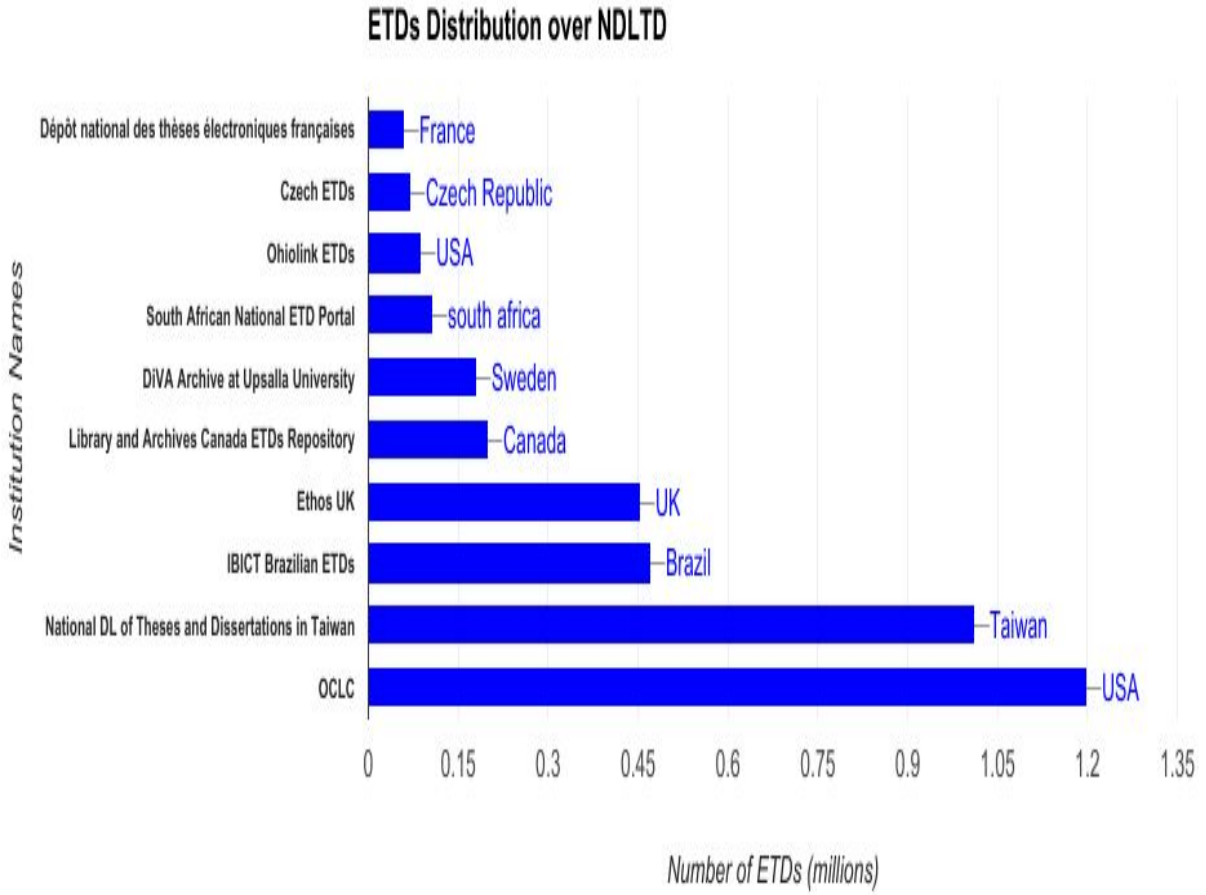


CiteSeerX



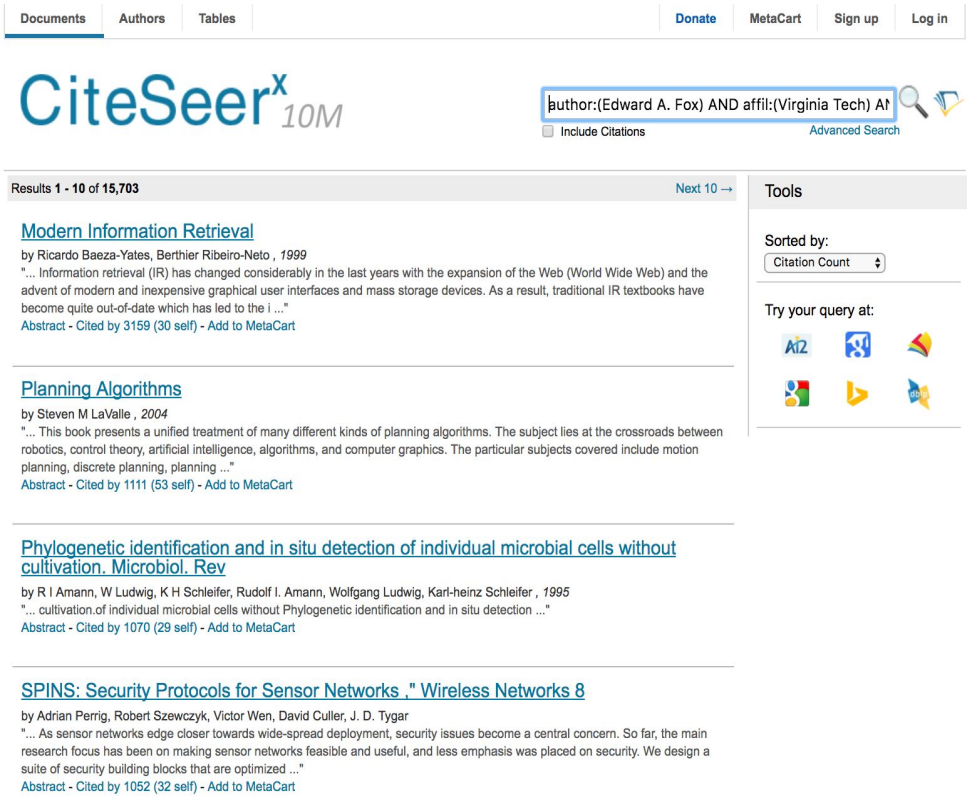
NDLTD

- ETDs Statistics
- Functional Limitations
  - Categorization Missing
  - No Full Text Access

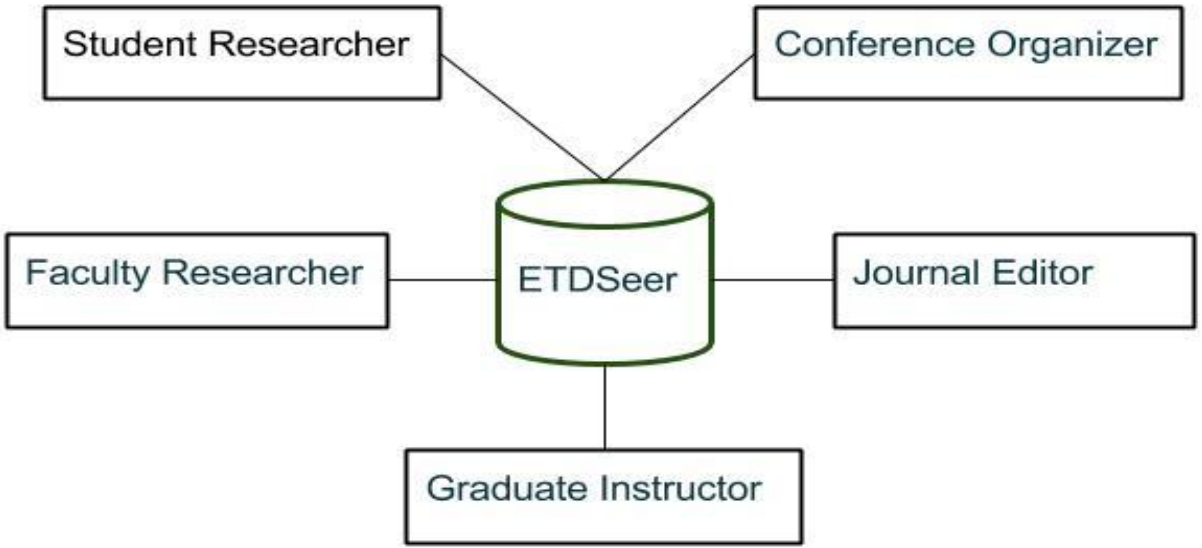


CiteSeerX

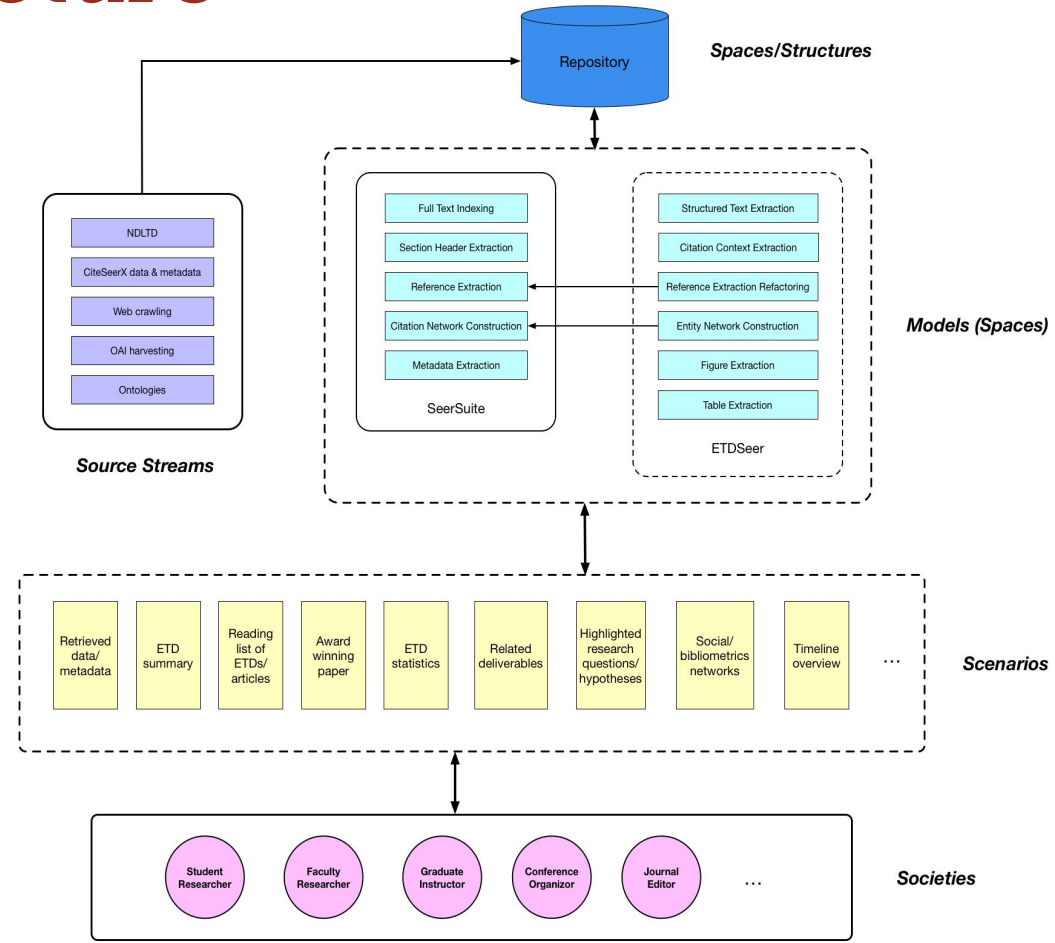
- Data: Make use of Metadata
- Technologies: SeerSuite tools
  - Automatic citation indexing
  - Automatic metadata extraction
  - Reference linking
  - Author disambiguation
  - Related documents
  - Full-text indexing



Stakeholders Overview



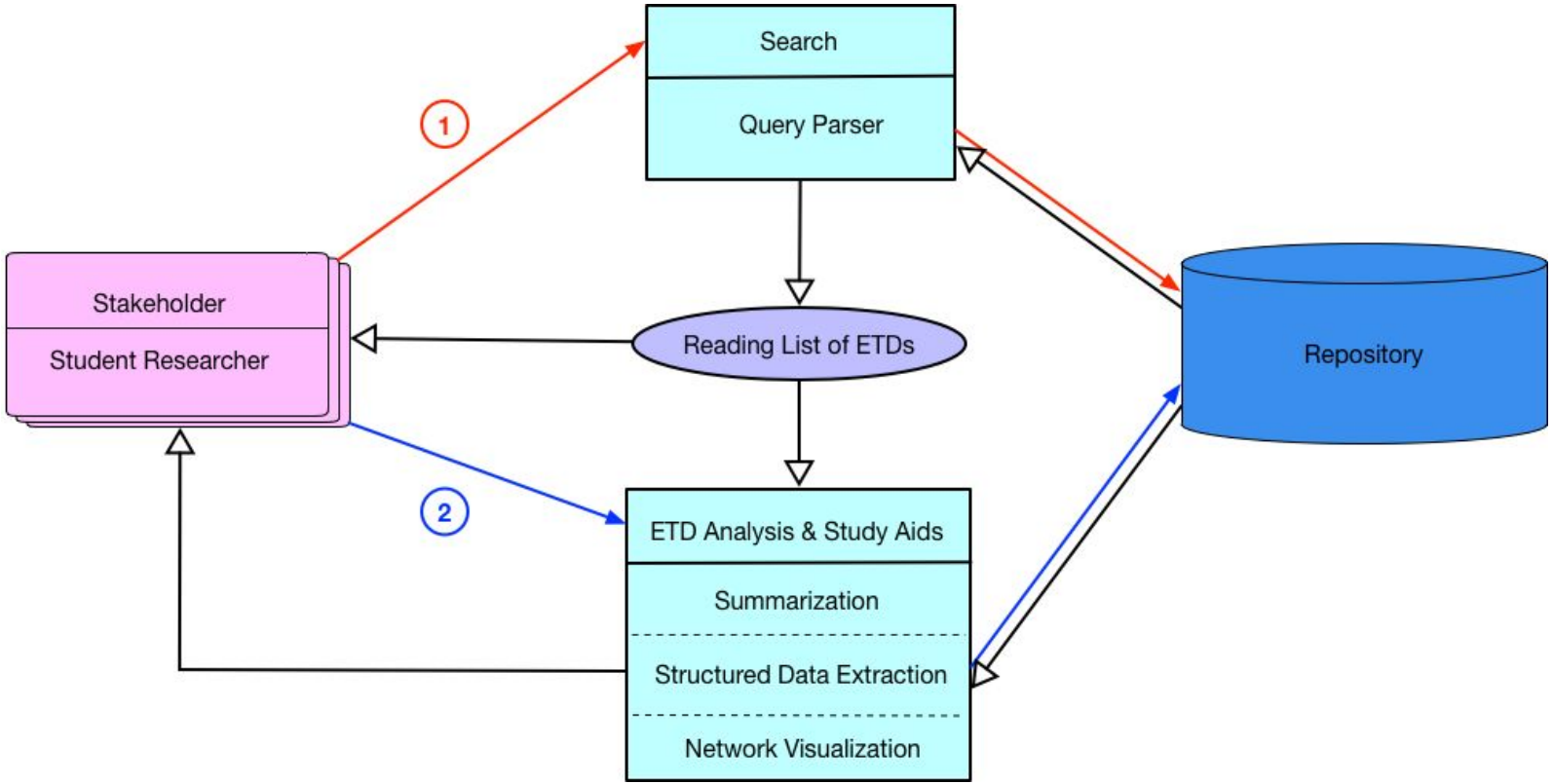
Architecture



Scenario 1 - Student Researcher

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

An Example Workflow

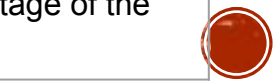


Other Scenarios

Stakeholder	Requirements	Expected Outputs
Faculty Researcher	Research problem exploration aid	• Synthesis of related ETDs • Proposed approaches/solutions • Future works summarization
Graduate Instructor	Graduate course syllabus formulation	• Draft with a hierarchical topical outline • Link to each topical entry with a reading list
	Advanced topic, lecture preparation	• Slides cover research questions/problems • Synthesis of provided potential solutions

Other Scenarios

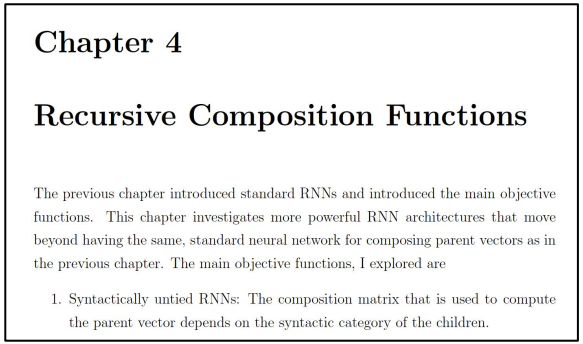
Stakeholder	Requirements	Expected Outputs
Conference Organizer	TPC member identification	• List of advisor research faculty names • Ranking table of advisors
	Potential participants identification	• Subgraph of ETD-derived citation graph • CSV file of author names, contact info.
Journal Editor	Peer-reviewer identification	Research interest-based reviewer list
	Content originality check	• Previous publications of the authors • Estimated percentage of the new content/work



Structured Data Extraction:

ETD Segmentation

- Heuristics-based strategy
  - Start with ‘Chapter’ or ‘CHAPTER’
  - Font size and style
- Deep learning approaches
  - Treat every two pages as one image
  - Manually label each image - chapter breaking point or not
  - Build a CNN model for classification

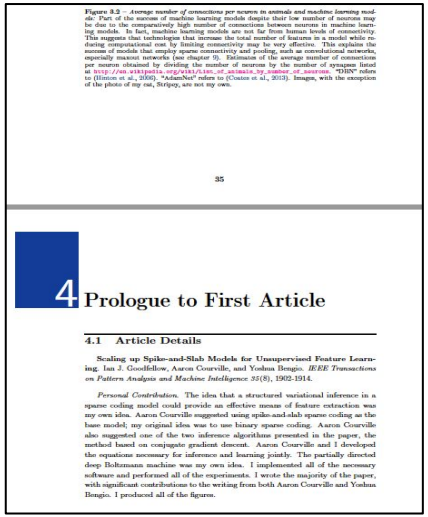


Chapter 4

Recursive Composition Functions

The previous chapter introduced standard RNNs and introduced the main objective functions. This chapter investigates more powerful RNN architectures that move beyond having the same, standard neural network for composing parent vectors as in the previous chapter. The main objective functions, I explored are

1. Syntactically untied RNNs: The composition matrix that is used to compute the parent vector depends on the syntactic category of the children.



35

4 Prologue to First Article

4.1 Article Details

Scaling up Spike-and-Slab Models for Unsupervised Feature Learning. Ian J. Goodfellow, Aaron Courville, and Yoshua Bengio. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(8), 1902-1914.

Personal Contribution. The idea that a structured variational inference in a sparse coding model could provide an effective means of feature extraction was my own idea. Aaron Courville suggested using spike-and-slab sparse coding as the base model; my original idea was to use binary sparse coding. Aaron Courville also suggested one of the two inference algorithms presented in the paper, the method based on conjugate gradient descent. Aaron Courville and I developed the equations necessary for inference and learning jointly. The partially directed deep Boltzmann machine was my own idea. I implemented all of the necessary software and performed all of the experiments. I wrote the majority of the paper, with significant contributions to the writing from both Aaron Courville and Yoshua Bengio. I produced all of the figures.

Structured Data Extraction:

Table & Figure Extraction

- **Extend work from TableSeer**
  - Table box detection
  - Table metadata extraction
  - Deal with styles of more disciplines
- **Figure Extraction**
  - Sagnik Ray Choudhury's work
  - CNN based approach - Mask R-CNN
  - AMT for figure labeling

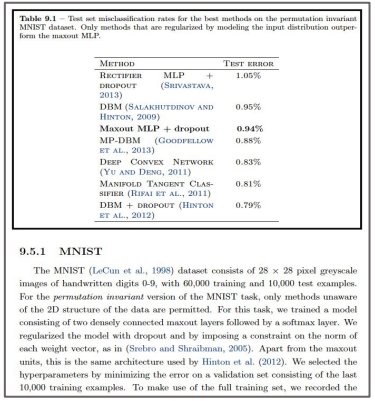
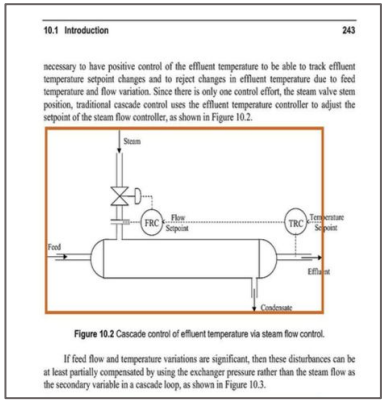


Table 9.1 — Test set misclassification rates for the best methods on the permutation invariant MNIST dataset. Only methods that are regularized by modeling the input distribution outperform the maxout MLP.

Method	Test error
Rectifier MLP + dropout (Srivastava, 2013)	1.05%
DBM (Salakhutdinov and Hinton, 2009)	0.95%
Maxout MLP + dropout	**0.94%**
MP-DBM (Goodfellow et al., 2013)	0.88%
Deep Convex Network (Yu and Deng, 2011)	0.83%
Manifold Tangent Classifier (Rifai et al., 2011)	0.81%
DBM + dropout (Hinton et al., 2012)	0.79%

9.5.1 MNIST

The MNIST (LeCun et al., 1998) dataset consists of 28 × 28 pixel greyscale images of handwritten digits 0-9, with 60,000 training and 10,000 test examples. For the *permutation invariant* version of the MNIST task, only methods unaware of the 2D structure of the data are permitted. For this task, we trained a model consisting of two densely connected maxout layers followed by a softmax layer. We regularized the model with dropout and by imposing a constraint on the norm of each weight vector, as in (Srebro and Shraibman, 2005). Apart from the maxout units, this is the same architecture used by Hinton et al. (2012). We selected the hyperparameters by minimizing the error on a validation set consisting of the last 10,000 training examples. To make use of the full training set, we recorded the

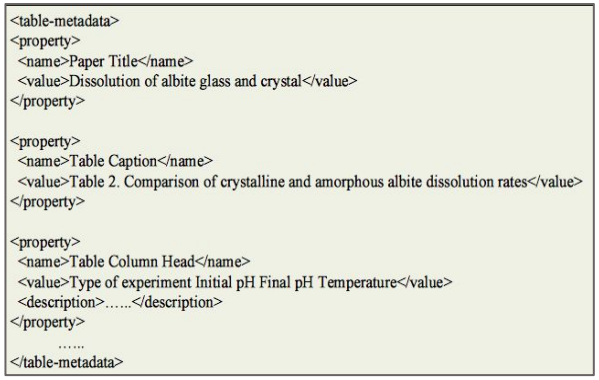


10.1 Introduction 243

necessary to have positive control of the effluent temperature to be able to track effluent temperature setpoint changes and to reject changes in effluent temperature due to feed temperature and flow variation. Since there is only one control effort, the steam valve stem position, traditional cascade control uses the effluent temperature controller to adjust the setpoint of the steam flow controller, as shown in Figure 10.2.

Figure 10.2 Cascade control of effluent temperature via steam flow control.

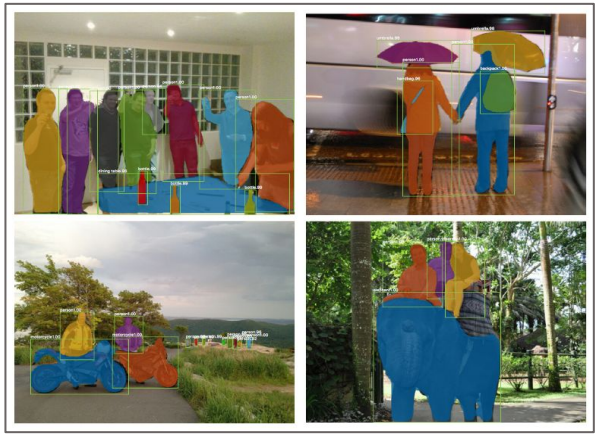
If feed flow and temperature variations are significant, then these disturbances can be at least partially compensated by using the exchanger pressure rather than the steam flow as the secondary variable in a cascade loop, as shown in Figure 10.3.



```
<table-metadata>
<property>
 <name>Paper Title</name>
 <value>Dissolution of albite glass and crystal</value>
</property>

<property>
 <name>Table Caption</name>
 <value>Table 2. Comparison of crystalline and amorphous albite dissolution rates</value>
</property>

<property>
 <name>Table Column Head</name>
 <value>Type of experiment Initial pH Final pH Temperature</value>
 <description>……</description>
</property>
        ……
</table-metadata>
```

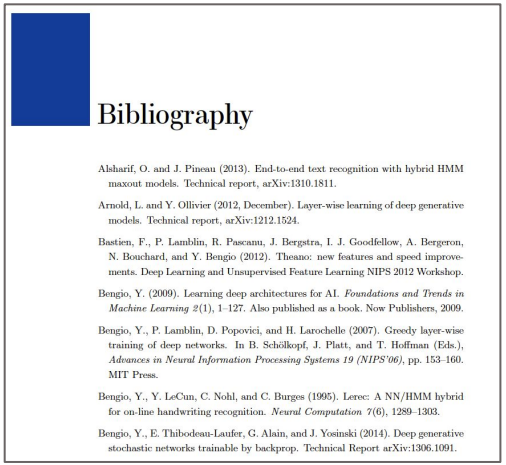


* K. He, et al., Mask R-CNN. *arXiv*, 2017.
* Y. Liu, et al., TableSeer, *JCDL* 2007.
* S. R. Choudhury, et al., An Architecture for information extraction from figures in DLs, *WWW*, 2015.

Structured Data Extraction:

Reference Extraction

- References that appear at the end
  - SeerSuite
- References that appear anywhere like footnotes
  - Deep learning for learning reference features
  - Classifier will be trained
- Represented in canonical format like BibTeX



Bibliography

Alsharif, O. and J. Pineau (2013). End-to-end text recognition with hybrid HMM maxout models. Technical report, arXiv:1310.1811.

Arnold, L. and Y. Ollivier (2012, December). Layer-wise learning of deep generative models. Technical report, arXiv:1212.1524.

Bastien, F., P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio (2012). Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning 2*(1), 1–127. Also published as a book. Now Publishers, 2009.

Bengio, Y., P. Lamblin, D. Popovici, and H. Larochelle (2007). Greedy layer-wise training of deep networks. In B. Schölkopf, J. Platt, and T. Hoffman (Eds.), *Advances in Neural Information Processing Systems 19 (NIPS'06)*, pp. 153–160. MIT Press.

Bengio, Y., Y. LeCun, C. Nohl, and C. Burges (1995). Lerec: A NN/HMM hybrid for on-line handwriting recognition. *Neural Computation* 7(6), 1289–1303.

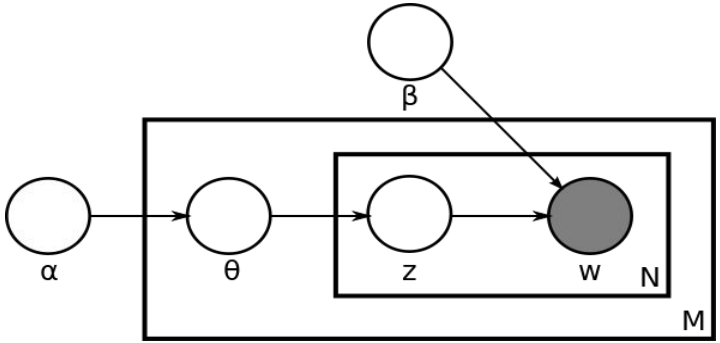
Bengio, Y., E. Thibodeau-Laufer, G. Alain, and J. Yosinski (2014). Deep generative stochastic networks trainable by backprop. Technical Report arXiv:1306.1091.

Donald E. Knuth's[1] Principal Principle is one of my favourite principles. One of my favourite books is *TEXbook*. Everybody should be rational.[2] Knuth said a lot of things. For instance, he said that everybody should be rational,[3] and he said that everybody should drive on the right side of the road.[4] Arnold van Gennep said that everybody should drive on the left, but otherwise van Gennep's[5] work agrees with Knuth.

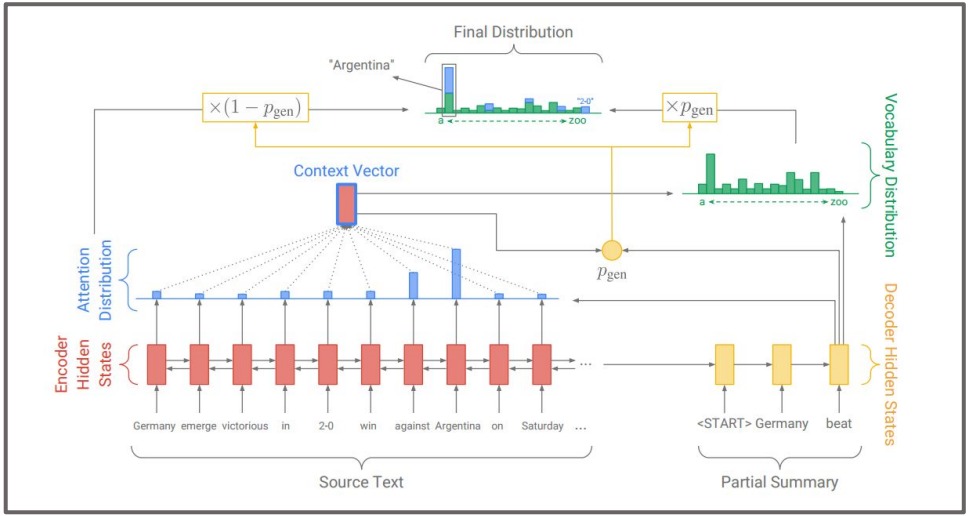
[1] Donald E. Knuth. *Computers & Typesetting*. Vol. A: *The TEXbook*. Reading, Mass.: Addison-Wesley, 1984.
[2] Knuth, *TEXbook*; Arnold van Gennep. *Les rites de passage*. Paris: Nourry, 1909.
[3] Knuth, *TEXbook*, p. 9.
[4] Donald E. Knuth. *Computers & Typesetting*. Vol. C: *The METAFONTbook*. Reading, Mass.: Addison-Wesley, 1986, pp. 10–15.
[5] Van Gennep, *Rites de passage*, p. 4.

Text Summarization

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

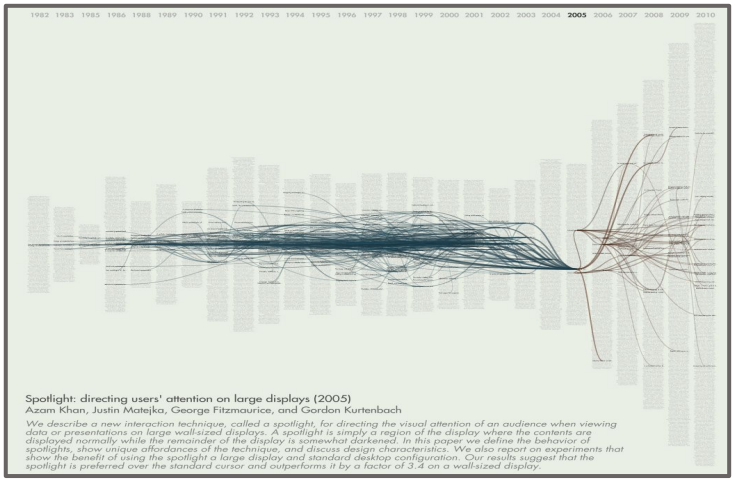


- Deep learning approach
  - Complete sentence
  - Attention model
  - Pointer & Generator

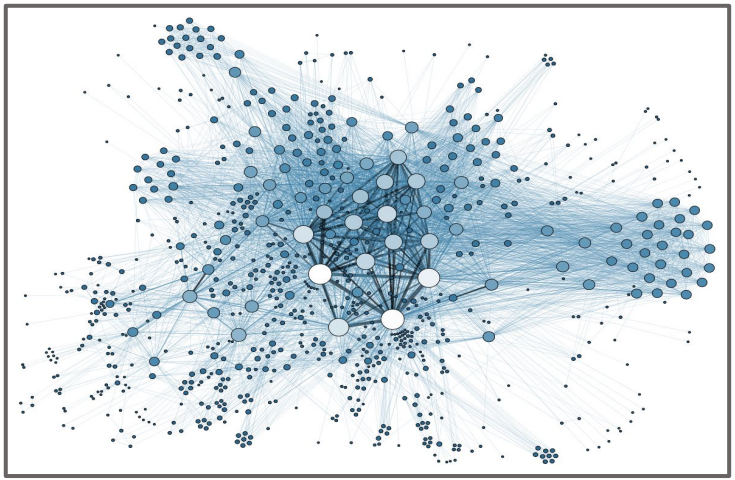


Network Visualization

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



- Social Network
  - Collaboration strength between research groups
  - Author social network



Working on…..

- User Study- based on the prospective Stakeholders
  - Questionnaires to reach broader audience
  - Interview (Focus group)
  - Hands on use and feedback
- Analysis of the collected data
- Presentation/Discussion of the result based on the analysis

Reference

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.

Cornelia Caragea, Jian Wu, Alina Ciobanu, Kyle Williams, Juan Fernandez-Ramirez,Hung-Hsuan Chen, Zhaohui Wu, and Lee Giles. CiteSeerX: A Scholarly Big Dataset, pages 311–322. Springer International Publishing, Cham, 2014.

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.

Complete list of references: https://goo.gl/QVgPQN

Thank you!!!