KDD’20 Tutorial: Scientific Text Mining and Knowledge Graphs

Meng Jiang
University of Notre Dame
Notre Dame, IN 46556, United States
mjiang2@nd.edu

Jingbo Shang
Univeristy of California, San Diego
La Jolla, CA 92037, United States
jshang@ucsd.edu

ABSTRACT

Unstructured scientific text, in various forms of textual artifacts, including manuscripts, publications, patents, and proposals, is used to store the tremendous wealth of knowledge discovered after weeks, months, and years, developing hypotheses, working in the lab or clinic, and analyzing results. A grand challenge on data mining research is to develop effective methods for transforming the scientific text into well-structured forms (e.g., ontology, taxonomy, knowledge graphs), so that machine intelligent systems can build on them for hypothesis generation and validation. In this tutorial, we provide a comprehensive overview on recent research and development in this direction. First, we introduce a series of text mining methods that extract phrases, entities, scientific concepts, relations, claims, and experimental evidence. Then we discuss methods that construct and learn from scientific knowledge graphs for accurate search, document classification, and exploratory analysis. Specifically, we focus on scalable, effective, weakly supervised methods that work on text in sciences (e.g., chemistry, biology).

1 BASIC INFORMATION

1.1 Target audience and prerequisites

This tutorial will be accessible to all data mining researchers, students, and practitioners who are interseted in text mining and knowledge graph technologies in scientific domains. No special prerequisite knowledge is needed to attend this tutorial.

1.2 Tutors and corresponding contact

Meng Jiang is an Assistant Professor in the Department of Computer Science and Engineering at the University of Notre Dame. His research interests include data mining, machine learning, and information extraction. He has published over 50 conference and journal papers of the topics. He has delivered *seven* tutorials in conferences such as KDD, SIGMOD, WWW, CIKM, ICDM, and SDM. He is the recipient of Notre Dame Global Gateway Faculty Award. (**In-person presenter** and **Corresponding tutor**)

Contact:

- Affiliation: Department of Computer Science and Engineering, University of Notre Dame
- Email: mjiang2@nd.edu
- Address: 384 Fitzpatrick Hall, Notre Dame, IN 46556, USA
- Phone: +1 (574) 631 7454

Jingbo Shang is an Assistant Professor at UC San Diego, jointly appointed by Computer Science Engineering (CSE) Department and Halıcıoğlu Data Science Institute (HDSI). His research focuses on mining and constructing structured knowledge from massive text corpora with minimum human effort. His research has been recognized with multiple prestigious awards, including Grand Prize of Yelp Dataset Challenge in 2015, Google PhD Fellowship in Structured Data and Database Management in 2017. He has rich experiences in delivering tutorials in major conferences, including KDD, WWW, SIGMOD, and VLDB. (**In-person presenter**)

Contact:

- Affiliation: CSE and HDSI, UC San Diego
- Email: jshang@ucsd.edu
- Address: Room 4104 at CSE Building, UC San Diego, La Jolla, CA 92037, USA
- Phone: +1 (217) 898 9691

2 TUTORIAL OUTLINE (3 HOURS)

1. Introduction and Overview (15 minutes)

2. Mining Structures and Learning Scientific Text (75 minutes)
   - 2.1. Automated phrase mining [10, 15]
   - 2.2. Scientific entity/concept recognition [16, 21, 22]
   - 2.3. Scientific relation extraction [4, 11, 12]
   - 2.4. Scientific language models [1, 9]
   - 2.5. Conditional statement extraction [6]
   - 2.6. Experimental evidence extraction [23, 24]

Presenters: Shang (2.1 – 2.4); Jiang (2.5, 2.6)

3. Constructing and Learning Scientific Knowledge Graphs (75 minutes)
   - 3.1. Ontology and hierarchy construction [2, 13, 25]
   - 3.2. Taxonomy generation and expansion [17, 18, 26]
   - 3.3. Knowledge graph construction [3, 7, 14]
   - 3.4. Learning for literature search and classification [5, 19]
   - 3.5. Learning for scientific text generation [8, 20]

Presenters: Shang (3.1, 3.2); Jiang (3.3 – 3.5)

4. Conclusions and Discussions (15 minutes)

3 OTHER INFORMATION

3.1 Compared to previous tutorials

There are three categories of related tutorials:

(I) Mining text in generic domains.

- Shang, et al., “Constructing and Mining Heterogeneous Information Networks from Massive Text,” KDD 2019 tutorial.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

KDD ’20, August 22-27, 2020, San Diego, CA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

- Shang, Jiang, et al., "Mining Entity-Relation-Attribute Structures from Massive Text Data," KDD 2017 tutorial.

Similarity: These related tutorials discussed text mining methods where the massive text refer to large scale of news articles, social media feeds, and product reviews.
Difference: Our tutorial will focus on recent development of text mining methods and knowledge graph algorithms in *scientific domains*. The specialized domains bring unique challenges such as lack of expert annotation and various kinds of expected structures (e.g., ontologies, taxonomies, conditions in scientific statements, and experimental evidence).

(II) Multidimensional analysis from text.

- Meng, et al., "TextCube: Automated Construction and Multidimensional Exploration," VLDB 2019 tutorial.
- Shang, et al., "Towards Multidimensional Analysis of Text Corpora," KDD 2018 tutorial.

Similarity: These related tutorials discussed text mining methods that project text data into multiple dimensions (e.g., person, location, event, sentiment). These methods used "cube" technologies instead of knowledge graph.
Difference: Our tutorial does not focus on multi-dimensional text analysis. We will focus on recent development of scientific text mining methods and scientific knowledge graph technologies.

(III) Knowledge graph in generic domains.

- Gao, et al., "Building a Large-scale, Accurate and Fresh Knowledge Graph," KDD 2018 tutorial.
- Shen, et al., "From Graph to Knowledge Graph: Mining Large-scale Heterogeneous Networks Using Spark," KDD 2019 hands-on tutorial.

Similarity: These related tutorials discussed knowledge graph construction and learning in generic domains.
Difference: Our tutorial will focus on recent development of knowledge graph algorithms in *scientific domains*. The specialized domains bring unique challenges as we have discussed above. Scientific knowledge graphs come from various forms of ontologies, taxonomies, and three-layer/heterogeneous graphs.

3.2 Strategies to encourage participation

If accepted, the tutorial notice will be disseminated through the presenters' homepage, their institution's webpage, their social media (e.g., Facebook, Twitter), and emails to communities of data mining, machine learning, specifically for those who are interested in the topics of scientific literature mining and knowledge graph. We will also pay attention to attendance of less representatives.

3.3 Potential societal impacts

This tutorial will provide a great opportunity for researchers in sciences (e.g., physics, chemistry, biology, psychology, social science) and those in data mining to talk to each other and find potential collaborations with each other. The technologies we will discuss in the tutorial need practice in real applications, real domains, and real problems with real needs. Their innovation and effectiveness will bring great power to facilitate the development of sciences.

3.4 Equipment

Equipment you will bring: We will bring laptop, pointer, and adapter. We will create a website for the tutorial and put all the slides on it before our presentation.

Equipment you will need: We will be well prepared and bring all equipment needed. We have no requirement to the conference.

Equipment attendees should bring: No required equipment for attendees. They can take notes using anything provided by the conference or brought by themselves (e.g., notebook, phone, laptop).

3.5 Video snippets

Videos of the presenter **Meng Jiang**, giving technical talks, can be found at http://www.meng-jiang.com/talks.html. The talks include two KDD 2014 oral presentations and one KDD 2016 oral presentation.

Videos of the presenter **Jingbo Shang**, giving technical talks, can be found at YouTube: https://youtu.be/YwLwzn6fw08?t=16008. This video is about his invited talk "Data-Driven Text Mining and Its Applications" at HDSI 2-year Anniversary.

REFERENCES

[1] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 3606–3611.

[2] Gene Ontology Consortium. 2017. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic acids research* 45, D1 (2017), D331–D338.

[3] Patrick Ernst, Amy Siu, and Gerhard Weikum. 2015. Knowlife: a versatile approach for constructing a large knowledge graph for biomedical sciences. *BMC bioinformatics* 16, 1 (2015), 157.

[4] Meng Jiang, Jingbo Shang, Taylor Cassidy, Xiang Ren, Lance M Kaplan, Timothy P Hanratty, and Jiawei Han. 2017. Metapad: Meta pattern discovery from massive text corpora. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 877–886.

[5] Tianwen Jiang, Zhihan Zhang, Tong Zhao, Bing Qin, Ting Liu, Nitesh V Chawla, and Meng Jiang. 2019. CTGA: Graph-based Biomedical Literature Search. In *Proceedings of the 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 395–400.

[6] Tianwen Jiang, Tong Zhao, Bing Qin, Ting Liu, Nitesh Chawla, and Meng Jiang. 2019. Multi-Input Multi-Output Sequence Labeling for Joint Extraction of Fact and Condition Tuples from Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 302–312.

[7] Tianwen Jiang, Tong Zhao, Bing Qin, Ting Liu, Nitesh V Chawla, and Meng Jiang. 2019. The Role of "Condition": A Novel Scientific Knowledge Graph Representation and Construction Model. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*. 1634–1642.

[8] Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text generation from knowledge graphs with graph transformers. In *NAACL*.

[9] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[10] Jialu Liu, Jingbo Shang, and Jiawei Han. 2017. Phrase mining from massive text and its applications. *Synthesis Lectures on Data Mining and Knowledge Discovery* 9, 1 (2017), 1–89.

[11] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multitask identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing and the 8th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

[12] Yi Luan, Mari Ostendorf, and Hannaneh Hajishirzi. 2017. Scientific information extraction with semi-supervised neural tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing and the 7th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

[13] Hoifung Poon and Pedro Domingos. 2010. Unsupervised ontology induction from text. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 296–305.

[14] Maya Rotmensch, Yoni Halpern, Abdulhakim Tlimat, Steven Horng, and David Sontag. 2017. Learning a health knowledge graph from electronic medical records. *Scientific reports* 7, 1 (2017), 1–11.

[15] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 30, 10 (2018), 1825–1837.

[16] Jingbo Shang, Liyuan Liu, Xiang Ren, Xiaotao Gu, Teng Ren, and Jiawei Han. 2018. Learning named entity tagger using domain-specific dictionary. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing and the 8th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

[17] Jingbo Shang, Xinyang Zhang, Liyuan Liu, Sha Li, and Jiawei Han. 2020. NetTaxo: Automated Topic Taxonomy Construction from Large-Scale Text-Rich Network. In *The Web Conference (TheWebConf)*.

[18] Jiaming Shen, Zhihong Shen, Chenyan Xiong, Chi Wang, Kuansan Wang, and Jiawei Han. 2020. TaxoExpan: Self-supervised Taxonomy Expansion with Position-Enhanced Graph Neural Network. In *The Web Conference (TheWebConf)*.

[19] Pingjie Tang, Meng Jiang, Ning Xia, Pitera J., Welser J., and Nitesh V Chawla. [n.d.]. Multi-label Patent Categorization with Non-local Attention-based Graph Convolutional Network. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. AAAI.

[20] Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019. Paperrobot: Incremental draft generation of scientific ideas. In *ACL*.

[21] Xuan Wang, Yu Zhang, Qi Li, Xiang Ren, Jingbo Shang, and Jiawei Han. 2019. Distantly Supervised Biomedical Named Entity Recognition with Dictionary Expansion. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 496–503.

[22] Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, and Jiawei Han. 2019. CrossWeigh: Training Named Entity Tagger from Imperfect Annotations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5157–5166.

[23] Wenhao Yu, Zongze Li, Qingkai Zeng, and Meng Jiang. 2019. Tablepedia: Automating pdf table reading in an experimental evidence exploration and analytic system. In *The World Wide Web Conference (WWW)*. 3615–3619.

[24] Wenhao Yu, Wei Peng, Yu Shu, Qingkai Zeng, and Meng Jiang. 2020. Experimental Evidence Extraction in Data Science with Hybrid Table Features and Ensemble Learning. In *The Web Conference (TheWebConf)*.

[25] Qingkai Zeng, Mengxia Yu, Wenhao Yu, Jinjun Xiong, Yiyu Shi, and Meng Jiang. 2019. Faceted hierarchy: A new graph type to organize scientific concepts and a construction method. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*. 140–150.

[26] Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian Sadler, Michelle Vanni, and Jiawei Han. 2018. Taxogen: Constructing topical concept taxonomy by adaptive term embedding and clustering. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*.