Bridging Text Data and Graph Data: Towards Semantics and Structure-aware Knowledge Discovery

Bowen Jin
University of Illinois at Urbana-Champaign
bowenj4@illinois.edu

Yu Zhang
University of Illinois at Urbana-Champaign
yuz9@illinois.edu

Sha Li
University of Illinois at Urbana-Champaign
shal2@illinois.edu

Jiawei Han
University of Illinois at Urbana-Champaign
hanj@illinois.edu

CCS CONCEPTS

• **Computing methodologies → Learning latent representations**; • **Information systems → Data mining**.

KEYWORDS

Pretrained Language Model, Graph Mining.

ACM Reference Format:
Bowen Jin, Yu Zhang, Sha Li, and Jiawei Han. 2024. Bridging Text Data and Graph Data: Towards Semantics and Structure-aware Knowledge Discovery. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24), March 4–8, 2024, Merida, Mexico.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3616855.3636450

1 MOTIVATION

Graphs and texts are two key modalities in data mining. In many cases, the data presents a mixture of the two modalities and the information is often complementary: in e-commerce data, the product-user graph and product descriptions capture different aspects of product features; in scientific literature, the citation graph, author metadata, and the paper content all contribute to modeling the paper impact.

However, the distinct properties of graph data and text data have led to the development of seemingly disparate methods. Graph neural networks are widely used for encoding graphs whereas language models trained with the Transformer architecture have become the mainstream for processing text. Early attempts are mainly a sequential application of the two models, which suffer from shallow semantic representations (by using frozen embeddings), or have scalability issues (joint training of two heterogeneous models). The community faces two critical questions:(1) how to represent text-rich networks (or network-enhanced text); and (2) is there a more organic way to integrate structure and semantic information?

In this tutorial, we review the recent developments in the representation and learning of text-rich networks [20], summarizing the best attempts to answer the questions above.

WSDM '24, March 4–8, 2024, Merida, Mexico.

ACM ISBN 979-8-4007-0371-3/24/03…$15.00
https://doi.org/10.1145/3616855.3636450

We start our tutorial by discussing how structure can be introduced into text data syntactically and semantically. This process could range from being local as in parsing a sentence into a dependency tree, and shallow as in creating a procedural graph with only temporal edges, to global and complex as in the construction of a knowledge graph. The type of data processing is application-specific and determines the semantics of the graph nodes and edges in the learning process. Next, we turn to introduce learning algorithms for text-rich networks (network-enhanced text). Based on the type of downstream task that we seek to optimize, we divide the methods into *network mining using language models* and *text mining with structure information*. Finally, we conclude the tutorial by counting the current successes, reflecting on lingering challenges, and pointing out directions for future research.

2 TUTORIAL OUTLINE

2.1 Introduction and Basic Concepts

We begin our tutorial by showing several examples of how graph and text are intertwined in real-life data (product networks [12], social networks [13], scientific literature networks [47], legal networks [48]) and then introduce the basic concepts and techniques for handling graph data and text data respectively.

For graph data, we will briefly introduce widely used techniques such as graph embeddings and graph neural networks (GNNs), and common graph-based tasks such as node classification, graph classification, and link prediction.

For text data, we will introduce pre-trained language models (including encoder-only models [7, 34, 37], encoder-decoder models [29, 40], and decoder-only models [3, 39]) and popular ways of utilizing pre-trained LMs, including fine-tuning, parameter-efficient tuning [14, 32], and in-context learning[3].

2.2 Enhancing text with graph structure

In this section, we cover a series of work on constructing graphs from text data, including sentence graphs, procedural graphs, and reasoning graphs.

2.2.1 Sentence-level graphs. Sentence-level graphs use words or concepts as nodes. Examples include dependency graphs, constituency graphs, AMR graphs, and OpenIE graphs. Graphs in this category usually have mature toolkits (Stanza, Spacy, StructBART) and we will demonstrate how the same input can be transformed into different graphs using the tools.

Sentence-level graphs are useful for fine-grained classification and extraction, such as their application in aspect-based sentiment analysis [56]. Ssentence-level graphs can also be used to decompose model outputs and perform fine-grained evaluation [41].

2.2.2 Procedure and event graphs. Procedural graphs use actions or steps as nodes and edges between steps denote their temporal order and logical dependency. The steps are represented by short phrases or sentences. In comparison, event graphs have structured events as nodes and there are multiple types of edges, including temporal edges, hierarchical edges, and causal edges.

If the edges are not observed, we show that it is possible to utilize the unsupervised alignment between steps and video transcripts [35, 69] to obtain the edges.

Once we collect many instances (imperfect observations) of the graph, we can then learn the graph structure by training a graph path model [69] or a graph generation model [31].

2.2.3 Belief and reasoning graphs. Graphs can be used to track the state of the world and the individual beliefs of characters [1, 43]. Compared to end-to-end approaches, this intermediate graph representation can provide a more focused view of the parts of the input that are relevant to the target object or person. Graph representations also help the model generalize to unseen configurations and support higher-order reasoning.

Graphs can also be used to represent explanations or reasoning behind answers. Instead of generating a single natural language explanation, entailment trees [6] and explanation graphs [42] provide a more structured way of manifesting the reasoning process of the model. By organizing the "thoughts" or intermediate reasoning steps of a language model into a graph [2, 52], we can observe significant improvements in the problem-solving ability of language models.

2.3 Knowledge graph construction

We will first introduce fundamental tasks of extracting phrases and named entities with distant supervision. Then, we will cover tasks that extract relations and structures connecting entities, such as taxonomy construction and knowledge graph construction for building a knowledge-preserving hierarchical structure.

2.3.1 Named entity recognition and entity typing. We will cover distantly supervised [30, 33] and few-shot [16, 17] named entity recognition methods, which aim to locate and classify named entities into pre-defined categories.

2.3.2 Relation and event extraction . Relation extraction identifies relations between named entities in text and helps build knowledge graphs linking multiple entities and their properties. We will cover recent studies on open-domain relation extraction [49, 68] and event extraction [19].

2.3.3 Coreference resolution and knowledge graph construction. After extracting entities, relations, and events, coreference resolution is needed to merge different mentions of the same entity/event and stitch the relations into a knowledge graph. We introduce some widely adopted approaches [26, 28] as well as some recent advances on multi-document [4] and long-document coreference resolution [45].

Finally, we show examples of how multiple information extraction components can be integrated for knowledge graph construction [9, 18].

2.4 Network mining with language models

In this section, we will emphasize how language models can mine networks with rich textual information (*i.e.*, text-rich networks). We will first introduce how graph neural networks are adopted on such networks. Then, we will discuss representation learning methods on networks with pretrained language models and how to pretrain language models with both semantic information and structure information.

2.4.1 Mining text-rich networks with graph neural networks. We will cover basic graph neural network (GNN) methods such as GCN [27], GraphSAGE [11] and GAT [46]. Then we will discuss GNN methods that encode semantic information together with structure information including TextGCN [51] and methods which proposes to refine the network with text information including BiTe-GCN [25] and AS-GCN [55].

2.4.2 Representation learning with language model on text-rich networks. We will first present language model architecture which can be adopted for representation learning on homogeneous text-rich networks where nodes [50] or edges [23] are associated with text information. We will then cover language model methods for representation learning on heterogeneous text-rich networks [22, 24].

2.4.3 Language model pretraining on text-rich networks. We will first briefly discuss basic language model pretraining strategies [7, 34]. Then, we will cover how to design structure-inductive strategies to better pretrain language models given a network of interests [21, 54], as well as its application on social media domain [59].

2.5 Text mining with structure information

In this section, we introduce how to leverage structure information (e.g., word-word co-occurrence graphs, metadata, citation links, knowledge graphs) for text mining tasks, which is dual to the topic introduced in Section 2.4. We will cover a wide variety of text mining tasks such as text classification, literature search, and question answering.

2.5.1 Graph-based/metadata-enhanced text classification. We start with methods using graph structures inside text (e.g., word-word co-occurrences, entity-document relationships, and the hierarchical structure of sections, subsections, and paragraphs) to enhance text classification. Related studies include the fully supervised HyperGAT [8], the semi-supervised HGAT [15], and the weakly supervised ClassKG [57] and FUTEX [63]. Then, we cover methods utilizing external metadata information (e.g., venues and authors of academic papers, users and products of e-commerce reviews) to construct graphs. Such metadata nodes and their combinations serve as additional signals to indicate categories. Related studies include the fully supervised MATCH [66], the semi-supervised MetaCat [65] and LTRN [60], as well as the weakly supervised META [36], MotifClass [62], and MICoL [67]. We will also introduce observations from two comprehensive benchmarking studies [10, 64].

2.5.2 *Citation-enhanced scientific literature understanding.*

Citation links contain rich semantic information to complement scientific documents. We will cover a series of studies on leveraging citations to enhance scientific language model pre-training. Earlier models such as SPECTER [5] and SciNCL [38] propose a citation-based contrastive pre-training paradigm, and they are evaluated on classification and recommendation tasks; more recent models such as SPECTER 2.0 [44] and SciMult [61] devise multi-task pre-training frameworks, which are evaluated on more diverse tasks such as literature search.

2.5.3 *Knowledge graph-enhanced question answering.*

Question answering is a challenging task that may require complex reasoning and external knowledge. We will introduce recent approaches, including GreaseLM [58], and DRAGON [53], that fuse contextualized language models and knowledge graphs during pre-training for commonsense reasoning and question answering.

2.6 Towards an Integrated Semantics and Structure Mining Paradigm

We have introduced graph construction from text, network mining with language models, and text mining with graph structure information. Such processing pipelines leave room for various kinds of deeper study on each component. Advanced methods can be further developed to index, organize, structure, and analyze text data and graph data and contribute to further knowledge discovery. Following this way, an integrated information process paradigm can be developed for organizing, manipulating, processing, and analyzing such integrated text and graph data for downstream applications. We will also outline our vision and some ongoing studies including how large foundation models can impact this line of work, as a conclusion of this tutorial.

3 FORMAT & SCHEDULE

The tutorial will be presented in **3 hours**, with 2 consecutive lecture-style sessions and a 15-minute break in between. The detailed schedule follows that described in Section 2.

- Introduction and basic concepts [15 mins, Jiawei Han]
- Enhancing text with graph structure [40 mins, Sha Li]
- Knowledge graph construction [40 mins, Jiawei Han]
- Break [15 mins]
- Network mining with language models [40 mins, Bowen Jin]
- Text mining with structure information [40 mins, Yu Zhang]
- Challenges and future work [10 mins, Jiawei Han]
- Q&A session [10 mins]

4 TUTORIAL MATERIAL

We will provide attendees with a website (https://peterjin.me/tutorials/wsdm24.html) and upload our tutorial materials (outline, slides, references, and software links) there.

5 PREVIOUS RELATED TUTORIALS

The following is a list of related tutorials with overlapped authors delivered at major international conferences in recent years:

(1) Xiang Ren, Meng Jiang, Jingbo Shang, and Jiawei Han, "*Constructing Structured Information Networks from Massive Text Corpora*" (WWW'17)
(2) Xiang Ren, Meng Jiang, Jingbo Shang, and Jiawei Han, "*Building Structured Databases of Factual Knowledge from Massive Text Corpora*" (SIGMOD'17)
(3) Jingbo Shang, Jiaming Shen, Liyuan Liu, and Jiawei Han, "*Constructing and Mining Heterogeneous Information Networks from Massive Text*" (KDD'19)

Differences from Previous Tutorials: Our new WSDM'24 tutorial proposal includes many pieces of recently published work after 2019 related to text mining and graph mining. Our focus will be different from previous versions (mainly focus on network construction), adding more content on mining network with pretrained language models and mining text with network structure information.

ACKNOWLEDGMENTS

Research was supported in part by US DARPA KAIROS Program No. FA8750-19-2-1004 and INCAS Program No. HR001121C0165, National Science Foundation IIS-19-56151, IIS-17-41317, and IIS 17-04532, and the Molecule Maker Lab Institute: An AI Research Institutes program supported by NSF under Award No. 2019897, and the Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) by NSF under Award No. 2118329. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and do not necessarily represent the views, either expressed or implied, of DARPA or the U.S. Government.

REFERENCES

[1] Ashutosh Adhikari, Xingdi Yuan, Marc-Alexandre Côté, Mikuláš Zelinka, Marc-Antoine Rondeau, Romain Laroche, Pascal Poupart, Jian Tang, Adam Trischler, and Will Hamilton. 2020. Learning dynamic belief graphs to generalize on text-based games. In *NeurIPS'20.*
[2] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. 2023. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687* (2023).
[3] Tom B. Brown, Benjamin Mann, and Nick Ryder et al. 2020. Language Models are Few-Shot Learners. In *NeurIPS'20.*
[4] Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021. Cross-document Coreference Resolution over Predicted Mentions. In *Findings ACL'21.*
[5] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL'20.*
[6] Bhavana Dalvi, Peter Jansen, Oyvind Tafjord, Zhengnan Xie, Hannah Smith, Leighanna Pipatanangkura, and Peter Clark. 2021. Explaining Answers with Entailment Trees. In *EMNLP'21.*
[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT'19.*
[8] Kaize Ding, Jianling Wang, Jundong Li, Dingcheng Li, and Huan Liu. 2020. Be More with Less: Hypergraph Attention Networks for Inductive Text Classification. In *EMNLP'20.*
[9] Xinya Du, Zixuan Zhang, and et al Li, Sha. 2022. RESIN-11: Schema-guided Event Prediction for 11 Newsworthy Scenarios. In *NAACL'22, System Demonstrations.*
[10] Lukas Galke and Ansgar Scherp. 2022. Bag-of-Words vs. Graph vs. Sequence in Text Classification: Questioning the Necessity of Text-Graphs and the Surprising Strength of a Wide MLP. In *ACL'22.*
[11] Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *NIPS'17.*
[12] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW'16.*

[13] Itai Himelboim, Marc A Smith, Lee Rainie, Ben Shneiderman, and Camila Espina. 2017. Classifying Twitter topic-networks using social network analysis. *Social Media + Society* (2017).

[14] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *ICML'19*.

[15] Linmei Hu, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. 2019. Heterogeneous graph attention networks for semi-supervised short text classification. In *EMNLP'19*.

[16] Jiaxin Huang, Chunyuan Li, Krishan Subudhi, Damien Jose, Shobana Balakrishnan, Weizhu Chen, Baolin Peng, Jianfeng Gao, and Jiawei Han. 2021. Few-Shot Named Entity Recognition: An Empirical Baseline Study. In *EMNLP'21*.

[17] Jiaxin Huang, Yu Meng, and Jiawei Han. 2022. Few-Shot Fine-Grained Entity Typing with Automatic Label Interpretation and Instance Generation. In *KDD'22*.

[18] Pengcheng Jiang, Shivam Agarwal, Bowen Jin, Xuan Wang, Jimeng Sun, and Jiawei Han. 2023. Text-Augmented Open Knowledge Graph Completion via Pre-Trained Language Models. *arXiv preprint arXiv:2305.15597* (2023).

[19] Yizhu Jiao, Sha Li, Yiqing Xie, Ming Zhong, Heng Ji, and Jiawei Han. 2022. Open-Vocabulary Argument Role Prediction for Event Extraction. In *EMNLP'22*.

[20] Bowen Jin, Gang Liu, Chi Han, Meng Jiang, Heng Ji, and Jiawei Han. 2023. Large Language Models on Graphs: A Comprehensive Survey. *arXiv preprint arXiv:2312.02783* (2023).

[21] Bowen Jin, Wentao Zhang, Yu Zhang, Yu Meng, Xinyang Zhang, Qi Zhu, and Jiawei Han. 2023. Patton: Language Model Pretraining on Text-Rich Networks. In *ACL'23*.

[22] Bowen Jin, Wentao Zhang, Yu Zhang, Yu Meng, Han Zhao, and Jiawei Han. 2023. Learning Multiplex Embeddings on Text-rich Networks with One Text Encoder. *arXiv preprint arXiv:2310.06684* (2023).

[23] Bowen Jin, Yu Zhang, Yu Meng, and Jiawei Han. 2023. Edgeformers: Graph-Empowered Transformers for Representation Learning on Textual-Edge Networks. In *ICLR'23*.

[24] Bowen Jin, Yu Zhang, Qi Zhu, and Jiawei Han. 2023. Heterformer: Transformer-based deep node representation learning on heterogeneous text-rich networks. In *KDD'23*.

[25] Di Jin, Xiangchen Song, Zhizhi Yu, Ziyang Liu, Heling Zhang, Zhaomeng Cheng, and Jiawei Han. 2021. Bite-gcn: A new gcn architecture via bidirectional convolution of topology and features on text-rich networks. In *WSDM'21*.

[26] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *TACL* (2020).

[27] Thomas N Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR'16*.

[28] Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference Resolution without Span Representations. In *ACL'21*.

[29] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *ACL'20*.

[30] Bangzheng Li, Wenpeng Yin, and Muhao Chen. 2022. Ultra-fine Entity Typing with Indirect Supervision from Natural Language Inference. *TACL* (2022).

[31] Manling Li, Sha Li, Zhenhailong Wang, Lifu Huang, Kyunghyun Cho, Heng Ji, Jiawei Han, and Clare Voss. 2021. The Future is not One-dimensional: Complex Event Schema Induction by Graph Modeling for Event Prediction. In *EMNLP'21*.

[32] Xiang Lisa Li and Percy Liang. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *ACL'21*.

[33] Chen Liang, Yue Yu, Haoming Jiang, Siawpeng Er, Ruijia Wang, Tuo Zhao, and Chao Zhang. 2020. BOND: BERT-Assisted Open-Domain Named Entity Recognition with Distant Supervision. In *KDD'20*.

[34] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[35] Lajanugen Logeswaran, Sungryull Sohn, Yunseok Jang, Moontae Lee, and Honglak Lee. 2023. Unsupervised Task Graph Generation from Instructional Video Transcripts. In *Findings of ACL'23*.

[36] Dheeraj Mekala, Xinyang Zhang, and Jingbo Shang. 2020. META: Metadata-Empowered Weak Supervision for Text Classification. In *EMNLP'20*.

[37] Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. COCO-LM: Correcting and Contrasting Text Sequences for Language Model Pretraining. In *NeurIPS'21*.

[38] Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood Contrastive Learning for Scientific Document Representations with Citation Embeddings. In *EMNLP'22*.

[39] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. In *OpenAI blog*.

[40] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR* (2020).

[41] Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating Factuality in Summarization with Semantic Graph Representations. In *NAACL'22*.

[42] Swarnadeep Saha, Prateek Yadav, Lisa Bauer, and Mohit Bansal. 2021. ExplaGraphs: An Explanation Graph Generation Task for Structured Commonsense Reasoning. In *EMNLP'21*.

[43] Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. 2023. Minding Language Models' (Lack of) Theory of Mind: A Plug-and-Play Multi-Character Belief Tracker. In *ACL'23*.

[44] Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. SciRepEval: A Multi-Format Benchmark for Scientific Document Representations. *arXiv preprint arXiv:2211.13308* (2022).

[45] Raghuveer Thirukovalluru, Nicholas Monath, Kumar Shridhar, Manzil Zaheer, Mrinmaya Sachan, and Andrew McCallum. 2021. Scaling Within Document Coreference to Long Texts. In *Findings of ACL'21*.

[46] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *ICLR'18*.

[47] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies* (2020).

[48] Ryan Whalen. 2016. Legal networks: The promises and challenges of legal network analysis. *Michigan State Law Review* (2016).

[49] Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. 2022. EIDER: Evidence-enhanced Document-level Relation Extraction. In *ACL'22*.

[50] Junhan Yang, Zheng Liu, Shitao Xiao, Chaozhuo Li, Defu Lian, Sanjay Agrawal, Amit Singh, Guangzhong Sun, and Xing Xie. 2021. GraphFormers: GNN-nested transformers for representation learning on textual graph. In *NeurIPS'21*.

[51] Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *AAAI'19*.

[52] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601* (2023).

[53] Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning, Percy Liang, and Jure Leskovec. 2022. Deep Bidirectional Language-Knowledge Graph Pretraining. In *NeurIPS'22*.

[54] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining Language Models with Document Links. In *ACL'22*.

[55] Zhizhi Yu, Di Jin, Ziyang Liu, Dongxiao He, Xiao Wang, Hanghang Tong, and Jiawei Han. 2021. AS-GCN: Adaptive semantic architecture of graph convolutional networks for text-rich networks. In *ICDM'21*.

[56] Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based Sentiment Classification with Aspect-specific Graph Convolutional Networks. In *EMNLP'19*.

[57] Lu Zhang, Jiandong Ding, Yi Xu, Yingyao Liu, and Shuigeng Zhou. 2021. Weakly-supervised Text Classification Based on Keyword Graph. In *EMNLP'21*.

[58] X Zhang, A Bosselut, M Yasunaga, H Ren, P Liang, C Manning, and J Leskovec. 2022. GreaseLM: Graph REASoning Enhanced Language Models for Question Answering. In *ICLR'22*.

[59] Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023. TwHIN-BERT: a socially-enriched pre-trained language model for multilingual Tweet representations. In *KDD'23*.

[60] Xinyang Zhang, Chenwei Zhang, Xin Luna Dong, Jingbo Shang, and Jiawei Han. 2021. Minimally-Supervised Structure-Rich Text Categorization via Learning on Text-Rich Networks. In *WWW'21*.

[61] Yu Zhang, Hao Cheng, Zhihong Shen, Xiaodong Liu, Ye-Yi Wang, and Jianfeng Gao. 2023. Pre-training Multi-task Contrastive Learning Models for Scientific Literature Understanding. *arXiv preprint arXiv:2305.14232* (2023).

[62] Yu Zhang, Shweta Garg, Yu Meng, Xiusi Chen, and Jiawei Han. 2022. Motifclass: Weakly supervised text classification with higher-order metadata information. In *WSDM'22*.

[63] Yu Zhang, Bowen Jin, Xiusi Chen, Yanzhen Shen, Yunyi Zhang, Yu Meng, and Jiawei Han. 2023. Weakly Supervised Multi-Label Classification of Full-Text Scientific Papers. In *KDD'23*.

[64] Yu Zhang, Bowen Jin, Qi Zhu, Yu Meng, and Jiawei Han. 2023. The Effect of Metadata on Scientific Literature Tagging: A Cross-Field Cross-Model Study. In *WWW'23*.

[65] Yu Zhang, Yu Meng, Jiaxin Huang, Frank F. Xu, Xuan Wang, and Jiawei Han. 2020. Minimally supervised categorization of text with metadata. In *SIGIR'20*.

[66] Yu Zhang, Zhihong Shen, Yuxiao Dong, Kuansan Wang, and Jiawei Han. 2021. MATCH: Metadata-Aware Text Classification in A Large Hierarchy. In *WWW'21*.

[67] Yu Zhang, Zhihong Shen, Chieh-Han Wu, Boya Xie, Junheng Hao, Ye-Yi Wang, Kuansan Wang, and Jiawei Han. 2022. Metadata-Induced Contrastive Learning for Zero-Shot Multi-Label Text Classification. In *WWW'22*.

[68] Sizhe Zhou, Suyu Ge, Jiaming Shen, and Jiawei Han. 2023. Corpus-Based Relation Extraction by Identifying and Refining Relation Patterns. In *ECML/PKDD'23*.

[69] Yu Zhou, Sha Li, Manling Li, Xudong Lin, Shih-Fu Chang, Mohit Bansal, and Heng Ji. 2023. Non-Sequential Graph Script Induction via Multimedia Grounding. In *ACL'23*.