Integrating Textual and Graph Data: Advancing Knowledge Discovery with Semantic and Structural Insights

Bowen Jin* Yu Zhang† Yunyi Zhang‡ Jiawei Han§

1 Basic Information

Graphs and textual data are both fundamental in the realm of data mining, each with its own unique features that often necessitate specialized modeling techniques. Traditionally, technologies for graph mining and text mining have been developed independently. However, in many real-world applications, these two data modalities frequently coexist and complement each other. For example, in e-commerce platforms, user-product interaction graphs and product descriptions provide different but complementary perspectives on product characteristics. Likewise, in scientific research, citation networks, author information, and the textual content of papers collectively shape the understanding of a paper's impact.

In this tutorial, we will focus on recent advances in graph mining techniques that harness the power of Large Language Models (LLMs), as well as improvements in text mining methods through the integration of graph structure information. Our goal is to present a cohesive framework illustrating how the synergy between graph and text modalities can lead to richer insights and more profound knowledge discovery. The key topics covered will include:

(1) An introduction to how graph and text data are interlinked in real-world applications and how models like graph neural networks and large language models are designed to capture signals from both modalities; (2) **Network mining with language models**: methods that utilize language models for representation learning on graphs and pretraining language models with graph data; (3) **Text mining with structural information**: approaches to text classification, document retrieval, and question answering, where graph structures serve as auxiliary information; (4) Progressing towards an integrated paradigm for mining both semantics and structural information in a unified framework.

This comprehensive exploration will provide insights into how combining graph structures and textual data can enhance the capabilities of modern machine-learning techniques. The tutorial will be presented in **2 hours**.

2 Target Audience

This tutorial is aimed at researchers and practitioners working in the domains of text mining, graph mining, information extraction, information retrieval, and knowledge discovery. Since text mining and graph mining are both frontier research directions, this tutorial will provide a unified view for future research in the SIAM data mining community. While participants with a solid background in these areas will gain the most from the content, we believe that the material will also serve as a valuable introduction for a broader audience, including those new to the field. The tutorial will provide an entry point to current advancements and key research challenges in this area, encouraging further exploration and engagement with the subject matter. We have designed the tutorial to be self-contained, ensuring that only a foundational understanding of data mining and machine learning concepts is required. This approach makes the tutorial accessible to participants at various levels of expertise while offering deeper insights for those already familiar with the core topics.

*University of Illinois at Urbana-Champaign. bowenj4@illinois.edu.
†University of Illinois at Urbana-Champaign. yuz9yuz@gmail.com
‡University of Illinois at Urbana-Champaign. yzhan238@illinois.edu
§University of Illinois at Urbana-Champaign. hanj@illinois.edu.

3 Biography

- **Bowen Jin** is a Ph.D. candidate in Computer Science from UIUC. His research focuses on mining text data and graph data. He received the Apple PhD Fellowship (2024) and the Yunni & Maxine Pao Memorial Fellowship (2024). He has numerous research publications at KDD, WWW, ICLR, ICML, and NeurIPs.

- **Yu Zhang** is an incoming Assistant Professor in Computer Science and Engineering at TAMU and a Ph.D. candidate at UIUC. His research focuses on structure-enhanced text mining for science. He received the UIUC Dissertation Completion Fellowship (2023) and the Yunni & Maxine Pao Memorial Fellowship (2022).

- **Yunyi Zhang** is a Ph.D. candidate in Computer Science from UIUC. His research focuses on weakly supervised text mining, text classification, and taxonomy construction. He has numerous research publications at KDD, EMNLP, AAAI, and WSDM.

- **Jiawei Han** is the Michael Aiken Chair Professor in Computer Science from UIUC. His research areas encompass data mining, text mining, data warehousing, and information network analysis, with over 1000 research publications. He is Fellow of ACM, Fellow of IEEE, and received numerous prominent awards, including ACM SIGKDD Innovation Award (2004) and IEEE Computer Society W. Wallace McDowell Award (2009).

4 Tutorial Outline

4.1 Introduction and Basic Concepts [15 mins] We begin our tutorial by presenting various examples of how graphs and text are interconnected in real-world data, such as product networks, social networks, scientific literature networks, and legal networks. Afterward, we introduce fundamental concepts and techniques for working with both graph and text data.

For graph data, we will cover widely used methods like graph embeddings and graph neural networks (GNNs), along with common tasks such as node classification, graph classification, and link prediction. For text data, we will explore pre-trained language models, including encoder-only models [4, 18, 2, 20], encoder-decoder models [16, 23], and decoder-only models [1, 22]. We will also discuss popular techniques for utilizing pre-trained language models, such as fine-tuning, parameter-efficient tuning [8, 17], and in-context learning [1].

4.2 Network mining with language models [45 mins] In this section, we will highlight how language models can extract valuable insights from networks that contain extensive textual information (i.e., text-attributed graphs). We will begin by explaining how graph neural networks are applied to these networks. Next, we will explore representation learning techniques that leverage pretrained language models on such networks, as well as methods for pretraining language models to incorporate both semantic and structural information.

Mining text-attributed graphs with graph neural networks. We will introduce fundamental graph neural network (GNN) methods, including GCN [15], GraphSAGE [7], and GAT [25]. Following that, we will explore GNN approaches that integrate semantic information with structural data, such as TextGCN [27], as well as techniques that enhance networks using textual information, including BiTe-GCN [14] and AS-GCN [31].

Representation learning with language model on text-attributed graphs. We will begin by introducing language model architectures designed for representation learning on homogeneous text-attributed graphs, where either nodes [26] or edges [12] are linked to textual information. Next, we will discuss language model approaches for representation learning on heterogeneous text-attributed graphs [13, 11].

Language model pretraining on text-attributed graphs. We will start by briefly discussing fundamental language model pretraining strategies. Following that, we will explore how to develop structure-inductive techniques to enhance the pretraining of language models for specific networks of interest [29, 10], and their application in the social media domain [34].

4.3 Text mining with structure information [45 mins] In this section, we discuss how to utilize structural information (such as word-word co-occurrence graphs, metadata, citation links, and knowledge graphs) for text mining tasks, complementing the topic covered in Section 4.2. We will address a range of text mining tasks, including text classification, literature search, and question answering.

Graph-based/metadata-enhanced text classification. We begin by exploring methods that use graph structures within text (such as word-word co-occurrences, entity-document relationships, and the hierarchical organization of sections, subsections, and paragraphs) to improve text classification. Notable studies in this area include the fully supervised HyperGAT [5], the semi-supervised HGAT [9], and the weakly supervised ClassKG [32] and FUTEX [38]. Next, we cover approaches that leverage external metadata (such as venues and authors in academic papers or users and products in e-commerce reviews) to construct graphs. These metadata nodes, and their combinations, provide additional signals for identifying categories. Relevant studies include the fully supervised MATCH [41], the semi-supervised MetaCat [40] and LTRN [35], and the weakly supervised META [19], MotifClass [37], and MICoL [42]. We will also discuss findings from two comprehensive benchmarking studies [6, 39].

Citation-enhanced scientific literature understanding. Citation links provide valuable semantic information that can enrich scientific documents. We will discuss a range of studies that leverage citations to improve scientific language model pre-training. Early models, such as SPECTER [3] and SciNCL [21], introduced citation-based contrastive pre-training methods and were evaluated on tasks like classification and recommendation. More recent models, including SPECTER 2.0 [24] and SciMult [36], have developed multi-task pre-training frameworks, which have been assessed on a broader array of tasks, such as literature search.

Knowledge graph-enhanced question answering. Question answering is a complex task that often demands advanced reasoning and the integration of external knowledge. We will explore recent approaches, such as QA-GNN [30], GreaseLM [33], and DRAGON [28], which combine contextualized language models with knowledge graphs during pre-training to enhance commonsense reasoning and question answering.

4.4 Towards an Integrated Semantics and Structure Mining Paradigm [15 mins] The pipelines of network mining with language models and text mining using graph structure information offer opportunities for deeper exploration of each component. More advanced methods can be developed to index, organize, structure, and analyze both text and graph data, contributing to further knowledge discovery. Building on this, an integrated information processing paradigm could emerge to organize, manipulate, process, and analyze combined text and graph data for various downstream applications. To conclude this tutorial, we will share our vision and discuss ongoing research, including how large foundation models may influence future developments in this field.

5 Similar 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. Jingbo Shang, Jiaming Shen, Liyuan Liu, and Jiawei Han, "*Constructing and Mining Heterogeneous Information Networks from Massive Text*" (KDD'19)
3. Bowen Jin, Yu Zhang, Sha Li, and Jiawei Han, "*Bridging Text Data and Graph Data: Towards Semantics and Structure-aware Knowledge Discovery*" (WSDM'24)

Compared with the first two tutorials, our tutorial will illustrate more about the connection between graph data and text data from a modern "large language model" perspective. Compared with the third tutorial, our tutorial will include more recent papers in the text & graph learning field.

References

[1] T. B. Brown, B. Mann, and N. R. et al, *Language models are few-shot learners*, in NeurIPS'20, 2020.

[2] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, *ELECTRA: Pre-training text encoders as discriminators rather than generators*, in ICLR'20, 2020.

[3] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. S. Weld, *Specter: Document-level representation learning using citation-informed transformers*, in ACL'20, 2020.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *BERT: Pre-training of deep bidirectional transformers for language understanding*, in NAACL-HLT'19, 2019.
[5] K. Ding, J. Wang, J. Li, D. Li, and H. Liu, *Be more with less: Hypergraph attention networks for inductive text classification*, in EMNLP'20, 2020.
[6] L. Galke and A. Scherp, *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, 2022.
[7] W. Hamilton, Z. Ying, and J. Leskovec, *Inductive representation learning on large graphs*, in NIPS'17, 2017.
[8] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, *Parameter-efficient transfer learning for nlp*, in ICML'19, 2019.
[9] L. Hu, T. Yang, C. Shi, H. Ji, and X. Li, *Heterogeneous graph attention networks for semi-supervised short text classification*, in EMNLP'19, 2019.
[10] B. Jin, W. Zhang, Y. Zhang, Y. Meng, X. Zhang, Q. Zhu, and J. Han, *Patton: Language model pretraining on text-rich networks*, in ACL'23, 2023.
[11] B. Jin, W. Zhang, Y. Zhang, Y. Meng, H. Zhao, and J. Han, *Learning multiplex embeddings on text-rich networks with one text encoder*, arXiv preprint arXiv:2310.06684, (2023).
[12] B. Jin, Y. Zhang, Y. Meng, and J. Han, *Edgeformers: Graph-empowered transformers for representation learning on textual-edge networks*, in ICLR'23, 2023.
[13] B. Jin, Y. Zhang, Q. Zhu, and J. Han, *Heterformer: Transformer-based deep node representation learning on heterogeneous text-rich networks*, in KDD'23, 2023.
[14] D. Jin, X. Song, Z. Yu, Z. Liu, H. Zhang, Z. Cheng, and J. Han, *Bite-gcn: A new gcn architecture via bidirectional convolution of topology and features on text-rich networks*, in WSDM'21, 2021.
[15] T. N. Kipf and M. Welling, *Semi-supervised classification with graph convolutional networks*, in ICLR'16, 2016.
[16] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, *BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*, in ACL'20, 2020.
[17] X. L. Li and P. Liang, *Prefix-tuning: Optimizing continuous prompts for generation*, in ACL'21, 2021.
[18] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, *RoBERTa: A robustly optimized bert pretraining approach*, arXiv preprint arXiv:1907.11692, (2019).
[19] D. Mekala, X. Zhang, and J. Shang, *Meta: Metadata-empowered weak supervision for text classification*, in EMNLP'20, 2020.
[20] Y. Meng, C. Xiong, P. Bajaj, S. Tiwary, P. Bennett, J. Han, and X. Song, *Coco-lm: Correcting and contrasting text sequences for language model pretraining*, in NeurIPS'21, 2021.
[21] M. Ostendorff, N. Rethmeier, I. Augenstein, B. Gipp, and G. Rehm, *Neighborhood contrastive learning for scientific document representations with citation embeddings*, in EMNLP'22, 2022.
[22] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, *Language models are unsupervised multitask learners*, in OpenAI blog, 2019.
[23] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, *Exploring the limits of transfer learning with a unified text-to-text transformer*, JMLR, (2020).
[24] A. Singh, M. D'Arcy, A. Cohan, D. Downey, and S. Feldman, *Scirepeval: A multi-format benchmark for scientific document representations*, arXiv preprint arXiv:2211.13308, (2022).
[25] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, *Graph attention networks*, in ICLR'18, 2018.
[26] J. Yang, Z. Liu, S. Xiao, C. Li, D. Lian, S. Agrawal, A. Singh, G. Sun, and X. Xie, *Graphformers: Gnn-nested transformers for representation learning on textual graph*, in NeurIPS'21, 2021.
[27] L. Yao, C. Mao, and Y. Luo, *Graph convolutional networks for text classification*, in AAAI'19, 2019.
[28] M. Yasunaga, A. Bosselut, H. Ren, X. Zhang, C. D. Manning, P. Liang, and J. Leskovec, *Deep bidirectional language-knowledge graph pretraining*, in NeurIPS'22, 2022.
[29] M. Yasunaga, J. Leskovec, and P. Liang, *Linkbert: Pretraining language models with document links*, in ACL'22, 2022.
[30] M. Yasunaga, H. Ren, A. Bosselut, P. Liang, and J. Leskovec, *Qa-gnn: Reasoning with language models and knowledge graphs for question answering*, in NAACL'21, 2021.
[31] Z. Yu, D. Jin, Z. Liu, D. He, X. Wang, H. Tong, and J. Han, *As-gcn: Adaptive semantic architecture of graph convolutional networks for text-rich networks*, in ICDM'21, 2021.
[32] L. Zhang, J. Ding, Y. Xu, Y. Liu, and S. Zhou, *Weakly-supervised text classification based on keyword graph*, in EMNLP'21, 2021.
[33] X. Zhang, A. Bosselut, M. Yasunaga, H. Ren, P. Liang, C. Manning, and J. Leskovec, *Greaselm: Graph reasoning enhanced language models for question answering*, in ICLR'22, 2022.
[34] X. Zhang, Y. Malkov, O. Florez, S. Park, B. McWilliams, J. Han, and A. El-Kishky, *Twhin-bert: a*

socially-enriched pre-trained language model for multilingual tweet representations, in KDD'23, 2023.
[35] X. Zhang, C. Zhang, X. L. Dong, J. Shang, and J. Han, *Minimally-supervised structure-rich text categorization via learning on text-rich networks*, in WWW'21, 2021.
[36] Y. Zhang, H. Cheng, Z. Shen, X. Liu, Y.-Y. Wang, and J. Gao, *Pre-training multi-task contrastive learning models for scientific literature understanding*, arXiv preprint arXiv:2305.14232, (2023).
[37] Y. Zhang, S. Garg, Y. Meng, X. Chen, and J. Han, *Motifclass: Weakly supervised text classification with higher-order metadata information*, in WSDM'22, 2022.
[38] Y. Zhang, B. Jin, X. Chen, Y. Shen, Y. Zhang, Y. Meng, and J. Han, *Weakly supervised multi-label classification of full-text scientific papers*, in KDD'23, 2023.
[39] Y. Zhang, B. Jin, Q. Zhu, Y. Meng, and J. Han, *The effect of metadata on scientific literature tagging: A cross-field cross-model study*, in WWW'23, 2023.
[40] Y. Zhang, Y. Meng, J. Huang, F. F. Xu, X. Wang, and J. Han, *Minimally supervised categorization of text with metadata*, in SIGIR'20, 2020.
[41] Y. Zhang, Z. Shen, Y. Dong, K. Wang, and J. Han, *Match: Metadata-aware text classification in a large hierarchy*, in WWW'21, 2021.
[42] Y. Zhang, Z. Shen, C.-H. Wu, B. Xie, J. Hao, Y.-Y. Wang, K. Wang, and J. Han, *Metadata-induced contrastive learning for zero-shot multi-label text classification*, in WWW'22, 2022.