Mining Structures from Massive Texts by Exploring the Power of Pre-trained Language Models

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

Yunyi Zhang
University of Illinois at Urbana-Champaign
yzhan238@illinois.edu

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

ABSTRACT

Technologies for handling massive structured or semi-structured data have been researched extensively in database communities. However, the real-world data are largely in the form of unstructured text, posing a great challenge to their management and analysis as well as their integration with semi-structured databases. Recent developments of deep learning methods and large pre-trained language models (PLMs) have revolutionized text mining and processing and shed new light on structuring massive text data and building a framework for integrated (i.e., structured and unstructured) data management and analysis.

In this tutorial, we will focus on the recently developed text mining approaches empowered by PLMs that can work without relying on heavy human annotations. We will present an organized picture of how a set of weakly supervised methods explore the power of PLMs to structure text data, with the following outline: (1) an introduction to pre-trained language models that serve as new tools for our tasks, (2) *mining topic structures*: unsupervised and seed-guided methods for topic discovery from massive text corpora, (3) *mining document structures*: weakly supervised methods for text classification, (4) *mining entity structures*: distantly supervised and weakly supervised methods for phrase mining, named entity recognition, taxonomy construction, and structured knowledge graph construction, and (5) towards an integrated information processing paradigm.

1 BACKGROUND, GOALS, AND DURATION

The massive text data available on the Web, social media, news, scientific literature, government reports, and other information sources contain rich knowledge that can potentially benefit a wide variety of information processing tasks, and they can be potentially structured and analyzed by extended database technologies. For example, one can conduct entity recognition and concept ontology construction on a large collection of scientific papers and extract the factual knowledge for knowledge base construction and subsequent analysis. How to effectively leverage the unstructured massive text data for downstream applications has remained an important and active research question for the past few decades. Recently, pre-trained language models (PLMs) such as BERT [6] have revolutionized the text mining field and brought new inspirations to structuring text data. To be specific, the following paradigm is usually adopted: pre-training neural architectures on large-scale text corpora obtained from the world knowledge (e.g., a combination of Wikipedia, books, scientific corpora, and web content), and then transferring their representations to task-specific data. By doing so, the knowledge encoded in the world corpora can be effectively leveraged to enhance downstream task performance significantly. However, the major challenge of such a paradigm is that fully supervised fine-tuning of PLMs usually requires abundant human annotations, which may require domain expertise and can be expensive and time-consuming to acquire in practice.

In this tutorial, we aim to introduce the recent developments in (1) language model pre-training that turns massive texts into contextualized text representations, and (2) weakly supervised methods that transfer pre-trained representations to various tasks for mining structures of topics, documents, and entities from massive texts. The materials introduced in our tutorial will greatly benefit researchers who work on text mining/natural language processing, data mining, and database systems, as well as practitioners who aim to obtain structured and actionable knowledge for targeted applications without access to abundant annotated data.

The tutorial will be presented in **3 hours**.

2 TUTORIAL OUTLINE

2.1 An Introduction to Pre-trained Language Models [40 mins]

PLMs effectively turn world-scale text corpora into text representations which can assist various kinds of downstream tasks for structuring a given corpus of text data.

2.1.1 Text Embedding and PLMs. We first provide an introduction to context-free embedding techniques, such as Word2Vec [35] and JoSE [28], and recent PLMs that learn contextualized representations based on the Transformer architecture, such as BERT [6], RoBERTa [24], ELECTRA [5], COCO-LM [30], and GPT [3, 37].

2.1.2 Common Usages of PLMs and Prompt-based Methods. We will then introduce common usages of PLMs in downstream tasks including standard fine-tuning, prompt-based fine-tuning [8, 38], lightweight tuning [12, 21], and zero-shot learning [29, 48] and inference [36].

2.2 Mining Topic Structures: Unsupervised and Seed-Guided Topic Discovery [35 mins]

Automatically mining a set of meaningful topics is one efficient way to digest large-scale text corpora. Traditional topic models (e.g., LDA [2] and SeededLDA [16]) are prominent tools for topic discovery. However, given the recent success of text representation learning and PLMs, it is reasonable to consider leveraging them to the quality of discovered topics.

2.2.1 Unsupervised Topic Discovery with PLMs. The high-quality text representations by PLMs can enhance the topic discovery process [1, 9] by forming more coherent topics. We will present

recent clustering methods [33, 43, 44] based on PLM embeddings for topic discovery.

2.2.2 Seed-guided, Discriminative Topic Discovery. Recent studies attempt to incorporate user guidance for specific topics in the topic discovery process to better fit a user's interests and needs. We will cover the following seed-guided topic discovery methods: CatE [27] takes a set of category names as guidance and trains text embeddings to capture term semantic similarity for topic discovery while enforcing distinctiveness across topics; JoSH [34] extends CatE into a hierarchical version by a user-provided taxonomy skeleton; KeyETM [11] extends the embedding-based topic model [7] to utilize seeds in the form of topic-level priors over the vocabulary; GTM [4] proposes a seed-guided topic-noise model for short texts; SeeTopic [57] leverages PLM-based text representations to deal with out-of-vocabulary seeds; SeedTopicMine [62] proposes to integrate multiple types of text representations (i.e., embeddings, PLMs, and topic-indicative sentences).

2.3 Mining Document Structures: Weakly Supervised Text Classification [35 mins]

Text classification aims to assign relevant labels to documents. Due to the cost and domain expertise needed for annotating sufficient, high-quality document-label pairs for supervision, some studies have been focusing on text classification with label names or a small set of training samples only. In this module, we introduce recent developments in weakly supervised text classification based on text embeddings and PLMs.

2.3.1 Flat Text Classification. We will first introduce the setting where the label space is flat. Related studies include ConWea [25], LOTClass [32], X-Class [47], and ClassKG [52] that explore PLMs as both general knowledge sources for understanding word semantics and strong representation learning methods for classification.

2.3.2 Hierarchical Text Classification. In many scenarios, categories form a tree/DAG-structured taxonomy [59]. We will present PCEM [49], a weakly-supervised hierarchical text classifier using a small set of training samples to perform efficient path prediction, as well as TaxoClass [40], a hierarchical multi-label text classifier using category names only, which employs PLMs for document-category similarity calculation.

2.3.3 Text Classification with Metadata. Documents on the Web are usually accompanied by metadata [55]. We will cover MetaCat [56], HIMECat [53], META[26], MotifClass [54], and HiGitClass [61] which jointly embed categories, text, and metadata into the same space and synthesize training samples based on the trained embeddings, as well as MICoL [60] which proposes a metadata-induced contrastive learning approach for zero-shot multi-label classification.

2.4 Mining Entity Structures: Taxonomy and Knowledge Base Construction [60 mins]

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.

2.4.1 Phrase Mining and Named Entity Recognition. The factual information in massive text corpora usually consists of entity mentions described by quality phrases. Incorporating external information from knowledge bases is a common practice in automated phrase mining (e.g., SegPhrase [23] and AutoPhrase [39]). We will present new phrase mining methods [10, 19] that use the self-attention mechanism in PLMs for phrase extraction. We will also cover distantly supervised [22, 31, 46] and few-shot [13, 14] named entity recognition methods, which aim to locate and classify named entities in unstructured text into pre-defined categories.

2.4.2 Taxonomy Construction. Taxonomy construction creates a hierarchy of "concept clusters" from massive corpora. Most existing taxonomies are constructed by human experts in a labor-intensive manner, not easily adaptable to changes in domains or users' interests. We will introduce CGExpan [58], FGExpan [50], TaxoExpan [41], CoRel [15], TaxoEnrich [17], and TaxoCom [20] to iteratively expand the user-given seed taxonomy and extract keywords for explaining each node.

2.4.3 Relation Extraction and Knowledge Graph Construction. 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 that explore the power of PLM for open-domain relation extraction [42, 51] and knowledge graph construction [18, 45].

2.5 Towards an Integrated Information Processing Paradigm [10 mins]

We have introduced a rich set of weakly supervised and PLM-enhanced approaches developed recently for automated structuring of massive text corpora. Such processing provides various kinds of rich semantic structures for subsequent developments, including quality phrases, typed entities, extracted relations, constructed knowledge graph fragments, classified documents, and typed heterogeneous information networks. Advanced methods can be further developed to index, organize, structure, and analyze such semantic primitives and integrate them with structured or semi-structured data in database systems. Following this way, an integrated information process paradigm can be developed for organizing, manipulating, processing, and analyzing such integrated, structured data for downstream applications. We will outline our vision and some ongoing studies in this direction, as a conclusion of this tutorial.

3 INTENDED AUDIENCE

Researchers and practitioners in the fields of database systems, data mining, text mining, natural language processing, information retrieval, and machine learning are targeted. While the audience with a good background in these areas would benefit most from this tutorial, we believe the material to be presented would give both the general audience and newcomers an introductory pointer to the current work and important research topics in this field, and inspire them to learn more. Our tutorial is designed as self-contained, so only preliminary knowledge about basic concepts in data mining, text mining, machine learning, and their applications is needed.

4 BIOGRAPHY

- **Yu Zhang** is a Ph.D. candidate in Computer Science from UIUC. His research focuses on weakly supervised text mining with structural information. He received the Yunni & Maxine Pao Memorial Fellowship (2022) and WWW Best Poster Award

Honorable Mention (2018). He has delivered tutorials in IEEE BigData'19, KDD'21, AAAI'22, and KDD'22.

- **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, WWW, WSDM, ACL, and EMNLP.
- **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). He delivered 50+ conference tutorials, including VLDB'19 and KDD'20-22 tutorials on a similar theme.

5 PREVIOUS RELATED TUTORIALS

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

(1) Yu Meng, Jiaxin Huang, Jingbo Shang, and Jiawei Han, "*TextCube: Automated Construction and Multidimensional Exploration*" (VLDB'19)

(2) Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han, "*On the Power of Pre-Trained Text Representations: Models and Applications in Text Mining*" (KDD'21)

(3) Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han, "*Pre-Trained Language Representations for Text Mining*" (AAAI'22)

(4) Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han, "*Adapting Pretrained Text Representations to Text Mining*" (KDD'22)

Differences from Previous Tutorials: Our new EDBT'23 tutorial proposal includes many pieces of recently published work in 2022 (and those to be published in 2023) with a focus on weakly supervised methods in structure mining and their big data applications. Parts of the contents have been presented in previous tutorials, with several more recent PLMs and their new applications added (e.g., zero-shot and few-shot learning for text mining, emergent properties, and techniques for large language models).

6 TUTORIAL MATERIAL

We will provide attendees with a website (https://yuzhimanhua.github.io/tutorials/edbt2023.html) and upload our tutorial materials (outline, slides, references, and software links) there.

ACKNOWLEDGMENTS

We thank anonymous reviewers for their valuable and insightful feedback. Research was supported in part by the IBM-Illinois Discovery Accelerator Institute, 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] Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. In *ACL'21*.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. In *JMLR*.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *NeurIPS'20*.

[4] Robert Churchill, Lisa Singh, Rebecca Ryan, and Pamela Davis-Kean. 2022. A Guided Topic-Noise Model for Short Texts. In *WWW'22*.

[5] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *ICLR'20*.

[6] 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*.

[7] Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic Modeling in Embedding Spaces. *TACL* (2020).

[8] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *ACL'21*.

[9] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[10] Xiaotao Gu, Zihan Wang, Zhenyu Bi, Yu Meng, Liyuan Liu, Jiawei Han, and Jingbo Shang. 2021. UCPhrase: Unsupervised Context-aware Quality Phrase Tagging. In *KDD'21*.

[11] Bahareh Harandizadeh, J. Hunter Priniski, and Fred Morstatter. 2022. Keyword Assisted Embedded Topic Model. In *WSDM'22*.

[12] 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*.

[13] 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*.

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

[15] Jiaxin Huang, Yiqing Xie, Yu Meng, Yunyi Zhang, and Jiawei Han. 2020. CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring. In *KDD'20*.

[16] Jagadeesh Jagarlamudi, Hal Daumé, and Raghavendra Udupa. 2012. Incorporating Lexical Priors into Topic Models. In *EACL'12*.

[17] Minhao Jiang, Xiangchen Song, Jieyu Zhang, and Jiawei Han. 2022. TaxoEnrich: Self-Supervised Taxonomy Completion via Structure-Semantic Representations. In *WWW'22*.

[18] 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*.

[19] Taeuk Kim, Jihun Choi, Daniel Edmiston, and Sang goo Lee. 2020. Are Pre-trained Language Models Aware of Phrases? Simple but Strong Baselines for Grammar Induction. In *ICLR'20*.

[20] Dongha Lee, Jiaming Shen, SeongKu Kang, Susik Yoon, Jiawei Han, and Hwanjo Yu. 2022. TaxoCom: Topic Taxonomy Completion with Hierarchical Discovery of Novel Topic Clusters. In *WWW'22*.

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

[22] 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*.
[23] Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han. 2015. Mining Quality Phrases from Massive Text Corpora. In *SIGMOD'15*.
[24] 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).
[25] Dheeraj Mekala and Jingbo Shang. 2020. Contextualized weak supervision for text classification. In *ACL'20*.
[26] Dheeraj Mekala, Xinyang Zhang, and Jingbo Shang. 2020. META: Metadata-Empowered Weak Supervision for Text Classification. In *EMNLP'20*.
[27] Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang, and Jiawei Han. 2020. Discriminative Topic Mining via Category-Name Guided Text Embedding. In *WWW'20*.
[28] Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance M. Kaplan, and Jiawei Han. 2019. Spherical Text Embedding. In *NeurIPS'19*.
[29] Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating Training Data with Language Models: Towards Zero-Shot Language Understanding. In *NeurIPS'22*.
[30] 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*.
[31] Yu Meng, Yunyi Zhang, Jiaxin Huang, Xuan Wang, Yu Zhang, Heng Ji, and Jiawei Han. 2021. Distantly-Supervised Named Entity Recognition with Noise-Robust Learning and Language Model Augmented Self-Training. In *EMNLP'21*.
[32] Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text Classification Using Label Names Only: A Language Model Self-Training Approach. In *EMNLP'20*.
[33] Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations. In *WWW'22*.
[34] Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, Chao Zhang, and Jiawei Han. 2020. Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding. In *KDD'20*.
[35] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NIPS'13*.
[36] Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. 2019. Language Models as Knowledge Bases?. In *EMNLP'19*.
[37] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. In *OpenAI blog*.
[38] Timo Schick and Hinrich Schütze. 2021. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *EACL'21*.
[39] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, and Jiawei Han. 2018. Automated Phrase Mining from Massive Text Corpora. *IEEE TKDE* (2018).
[40] Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren, and Jiawei Han. 2021. TaxoClass: Hierarchical Multi-Label Text Classification Using Only Class Names. In *NAACL-HLT'21*.
[41] 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 *WWW'20*.
[42] Jiaming Shen, Yunyi Zhang, Heng Ji, and Jiawei Han. 2021. Corpus-based Open-Domain Event Type Induction. In *EMNLP'21*.
[43] Suzanna Sia, Ayush Dalmia, and Sabrina J Mielke. 2020. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too!. In *EMNLP'20*.
[44] Laure Thompson and David Mimno. 2020. Topic modeling with contextualized word representation clusters. *arXiv preprint arXiv:2010.12626* (2020).
[45] Qingyun Wang, Manling Li, Xuan Wang, Nikolaus Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, Haoran Zhang, Weili Liu, et al. 2021. COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation. In *NAACL-HLT System Demonstrations'21*.
[46] Xuan Wang, Vivian Hu, Xiangchen Song, Shweta Garg, Jinfeng Xiao, and Jiawei Han. 2021. ChemNER: Fine-Grained Chemistry Named Entity Recognition with Ontology-guided Distant Supervision. In *EMNLP'21*.
[47] Zihan Wang, Dheeraj Mekala, and Jingbo Shang. 2021. X-Class: Text Classification with Extremely Weak Supervision. In *NAACL-HLT'21*.
[48] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models Are Zero-Shot Learners. In *ICLR'22*.
[49] Huiru Xiao, Xin Liu, and Yangqiu Song. 2019. Efficient path prediction for semi-supervised and weakly supervised hierarchical text classification. In *WWW'19*.
[50] Jinfeng Xiao, Mohab Elkaref, Nathan Herr, Geeth De Mel, and Jiawei Han. 2023. Taxonomy-Guided Fine-Grained Entity Set Expansion. In *SDM'23*.
[51] Yiqing Xie, Jiaming Shen, Sha Li, Yuning Mao, and Jiawei Han. 2022. EIDER: Evidence-enhanced Document-level Relation Extraction. In *ACL'22*.
[52] Lu Zhang, Jiandong Ding, Yi Xu, Yingyao Liu, and Shuigeng Zhou. 2021. Weakly-supervised Text Classification Based on Keyword Graph. In *EMNLP'21*.
[53] Yu Zhang, Xiusi Chen, Yu Meng, and Jiawei Han. 2021. Hierarchical Metadata-Aware Document Categorization under Weak Supervision. In *WSDM'21*.
[54] 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*.
[55] 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*.
[56] 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*.
[57] Yu Zhang, Yu Meng, Xuan Wang, Sheng Wang, and Jiawei Han. 2022. Seed-Guided Topic Discovery with Out-of-Vocabulary Seeds. In *NAACL'22*.
[58] Yunyi Zhang, Jiaming Shen, Jingbo Shang, and Jiawei Han. 2020. Empower Entity Set Expansion via Language Model Probing. In *ACL'20*.
[59] Yu Zhang, Zhihong Shen, Yuxiao Dong, Kuansan Wang, and Jiawei Han. 2021. MATCH: Metadata-Aware Text Classification in A Large Hierarchy. In *WWW'21*.
[60] 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*.
[61] Yu Zhang, Frank F. Xu, Sha Li, Yu Meng, Xuan Wang, Qi Li, and Jiawei Han. 2019. Higitclass: Keyword-driven hierarchical classification of github repositories. In *ICDM'19*.
[62] Yu Zhang, Yunyi Zhang, Martin Michalski, Yucheng Jiang, Yu Meng, and Jiawei Han. 2023. Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts. In *WSDM'23*.