Automated Taxonomy Discovery and Exploration

Jiaming Shen, Xiaotao Gu, Yu Meng, Jiawei Han
Department of Computer Science, University of Illinois at Urbana-Champaign, IL, USA
{*js2, xiaotao2, yumeng5, hanj*}*@illinois.edu*

***Abstract*—People nowadays are inundated with vast amounts of text data (*e.g.*, news articles, corporate reports, scientific papers, *etc.*). Turning massive text data into actionable knowledge is an essential research issue in data science. Based on our vision, it is highly beneficial to first structure raw text using taxonomies and then analyze structured text data in a more fine-grained and user-guided way.**

In this tutorial, we provide a comprehensive overview of recent research in this direction. We first show a series of methods to identify concept phrases from text corpora and then present methods to organize identified concepts into taxonomies. After that, we introduce techniques to automatically enrich an existing taxonomy and discuss how to explore taxonomies for different downstream applications. Finally, we demonstrate on real-world datasets from multiple domains how different taxonomies can be constructed based on different user tasks and how they can empower knowledge discovery from text data.

***Index Terms*—Phrase Mining, Taxonomy Construction, Taxonomy Enrichment, Weakly-supervised Text Classification**

I. Rationale: Why is this tutorial important?

In an era of information explosion, people are inundated with vast amounts of text data. Every day, there are thousands of scientific papers, tens of thousands of corporate reports, product reviews, and millions of social media posts produced and shared worldwide. Consequently, turning massive unstructured text data into actionable knowledge is an essential research issue in data science, which lays the foundation for realizing machine intelligence.

To bring this goal to reality, we vision that the taxonomy, which organizes concepts into hierarchical structures, will play a vital role. This is because the taxonomy can structure raw text in a versatile way and facilitate more fine-grained and semantic text analysis. So far, few studies in data mining, machine learning, and natural language processing communities have paid enough attention to exploring the power of leveraging taxonomy for discovering knowledge in text data. This tutorial is to bridge this gap and presents an overview of recent developments in this direction. Specifically, we will focus on effort-light approach, which relies less on human annotations. We will cover four major themes: (1) concept phrase mining, (2) automated taxonomy construction, (3) taxonomy enrichment, and (4) methods distilling knowledge from taxonomies to empower different downstream applications.

II. Content details

- Introduction (30min)
  - Motivations: Why constructing taxonomies and leveraging them to facilitate knowledge discovery from text data?
  - An overview of tasks related to taxonomy discovery such as concept phrase mining, automated taxonomy construction, and taxonomy enrichment.
  - An overview of applications that can benefit from knowledge in automatically discovered taxonomies.
- Phrase Mining (45min)
  - Why phrase mining and how to define properties of high-quality phrases?
  - Supervised phrase mining methods based on noun phrase chunking [5] and dependency parsing [6].
  - Weakly/Distantly supervised phrase mining methods [1], [3] utilizing external knowledge bases.
  - Unsupervised phrase mining methods exploring signals from topic model [4] and pre-trained language model [2].
  - System demos and software introduction:
    - A multilingual phrase mining system which integrates AutoPhrase [1] and TopMine [4] together and supports phrase mining in multiple languages (*e.g.*, English, Spanish, Chinese, Arabic, and Japanese).
    - A state-of-the-art phrase mining system [2] which leverages pre-trained language model to identify high-quality phrases and perform phrasal segmentation.
- Taxonomy Construction (45min)
  - How to represent a taxonomy based on different application needs?
  - Concept taxonomy construction
    - Two-step approaches: first extract hypernymy pairs and then organize extracted pairs into taxonomy structure [10], [11], [14]
    - End-to-end approaches: (1) supervised method based on structure prediction [15], and (2) weakly supervised method based on user seed guidance [7].
  - Topic taxonomy construction
    - Hierarchical topic modeling [13]
    - Hierarchical clustering [8]
    - Combining text data with network structure [9]
- Taxonomy Enrichment (45min)
  - What are different types of taxonomy enrichment?
  - WordNet enrichment method [17]
  - General taxonomy expansion methods:
    - Leveraging knowledge transfer techniques [18]
    - Modeling implicit taxonomic relation semantics [20]
    - Utilizing position-enhanced graph neural network [16]
    - Combining features from multiple sources [19]
  - General taxonomy modification methods:

    - ∗ Modifying existing taxonomic relations [22]
    - ∗ Generating new emerging concepts for direct taxonomy completion [21]
- Taxonomy Empowered Applications (45min)
  - Weakly-supervised text classification
    - ∗ One-level multi-class classification [25], [26]
    - ∗ Hierarchical multi-class classification [23], [24]
    - ∗ Hierarchical multi-label classification [27]
  - Query understanding and recommender system [28], [29]
- Summary and Future Directions (30min)
  - Summarizing presented principles and techniques
  - Discussing future research directions
  - Interacting with the audience and discussing how to discover and explore taxonomies based on their own data and applications.

III. Target audience and prerequisites

Potential participants can be researchers and practitioners in the fields of data mining, natural language processing, information retrieval, and machine learning. While the audience with a good background in the above areas would benefit most from this tutorial, we will design it to be self-contained so newcomers with basic knowledge of probability and high school mathematics can also benefit from those materials. We expect 50 to 100 participants and they can learn the current work and important research directions of phrase mining, taxonomy construction, taxonomy enrichment, and taxonomy-empowered applications.

IV. Related Conference Tutorials

The following are related tutorials with overlapping authors in recent years. Comparing to the previous tutorials, this tutorial covers and focuses on a lot of new materials. The remaining contents are new, as outlined below: (1) new methods on leveraging a pre-trained language model (LM) for better phrase mining and phrasal segmentation, (2) new taxonomy construction methods based on a pre-trained LM, (3) a complete new section on taxonomy enrichment, and (4) new weakly-supervised text classification methods and more applications empowered with taxonomy.

1) Yu Meng, Jiaxin Huang, and Jiawei Han. "Embedding-Driven Multi-Dimensional Topic Mining and Text Analysis". In *KDD* 2020.
2) Jingbo Shang, Jiaming Shen, Liyuan Liu, and Jiawei Han. "Constructing and Mining Heterogeneous Information Networks from Massive Text". In *KDD* 2019.
3) Jingbo Shang, Chao Zhang, Jiaming Shen, and Jiawei Han. "Towards Multidimensional Analysis of Text Corpora". In *KDD* 2018.

V. Tutor's short bio

We have four tutors and all of us will present this tutorial.

- **Jiaming Shen**, Ph.D. candidate, Computer Science, Univ. of Illinois at Urbana-Champaign (UIUC). His research focuses on unleashing hidden knowledge buried in unstructured text using taxonomy structures. He has been awarded several fellowships and scholarships, including a Brian Totty Graduate Fellowship and a Yunni & Maxine Pao Memorial Fellowship. Mr. Shen has delivered tutorials in KDD'18 and KDD'19.
- **Xiaotao Gu**, Ph.D. candidate, Computer Science, UIUC. His research focuses on automated knowledge mining from unstructured text data in real scenarios, *e.g.*, with limited human annotation and computation resource. His research work has been applied in several real-world systems, including Google News, Google Search, and the AMiner literature search and analytic platform.
- **Yu Meng**, Ph.D. candidate, Computer Science, UIUC. His research focuses on mining structured knowledge from massive text corpora with minimum human supervision. He received the Google Ph.D. Fellowship (2021) in Structured Data and Database Management. He has delivered tutorials in VLDB'19 and KDD'20.
- **Jiawei Han**, Michael Aiken Chair Professor, Computer Science, UIUC. His research areas encompass data mining, text mining, data warehousing, and information network analysis, with over 800 research publications. He is a 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 or keynote speeches (*e.g.*, SIGKDD 2019 tutorial and CIKM 2019 keynote).

VI. Proposed length of the tutorial

We plan for a 5 hours (full day) tutorial with 4 hours main lecture contents and 1 hour break.

References

[1] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, and Jiawei Han. "Automated Phrase Mining from Massive Text Corpora". In *TKDE* 2018.

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

[3] Bing Li, Xiaochun Yang, Bin Wang, and Wei Cui. "Efficiently Mining High Quality Phrases from Texts". In *AAAI* 2017.

[4] Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, and Jiawei Han. "Scalable Topical Phrase Mining from Text Corpora". In *VLDB* 2015.

[5] Florian Boudin. "PKE: An Open Source Python-based Keyphrase Extraction Toolkit". In *COLING (Demo)* 2016.

[6] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. "The Stanford CoreNLP Natural Language Processing Toolkit". In *ACL (Demo)* 55–60.

[7] Jiaming Shen, Zeqiu Wu, Dongming Lei, Chao Zhang, Xiang Ren, Michelle T. Vanni, Brian M. Sadler, and Jiawei Han. "HiExpan: Task-Guided Taxonomy Construction by Hierarchical Tree Expansion". In *KDD* 2018.

[8] Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian M. Sadler, Michelle T. Vanni, and Jiawei Han. "TaxoGen: Unsupervised Topic Taxonomy Construction by Adaptive Term Embedding and Clustering". In *KDD* 2018.

[9] Jingbo Shang, Xinyang Zhang, Liyuan Liu, Sha Li, and Jiawei Han. "NetTaxo: Automated Topic Taxonomy Construction from Text-Rich Network". In *WWW* 2020.

[10] Catherine Chen, Kevin Lin, and D. Klein. "Constructing Taxonomies from Pretrained Language Models". In *NAACL* 2021.

[11] Stephen Roller, Douwe Kiela, and Maximilian Nickel. "Hearst Patterns Revisited: Automatic Hypernym Detection from Large Text Corpora". In *ACL* 2018.

[12] Xueqing Liu, Yangqiu Song, Shixia Liu, and Haixun Wang. "Automatic taxonomy construction from keywords". In *KDD* 2012.

[13] David Blei, T. Griffiths, Michael I. Jordan, and J. Tenenbaum. "Hierarchical Topic Models and the Nested Chinese Restaurant Process". In *NIPS* 2003.

[14] Amit Gupta, R. Lebret, Hamza Harkous, and K. Aberer. "Taxonomy Induction Using Hypernym Subsequences". In *CIKM* 2017

[15] Mohit Bansal, David Burkett, Gerard de Melo, and D. Klein. "Structured Learning for Taxonomy Induction with Belief Propagation". In *ACL* 2014.

[16] Jiaming Shen, Zhihong Shen, Chenyan Xiong, Chi Wang, Kuansan Wang, and Jiawei Han. "TaxoExpan: Self-supervised Taxonomy Expansion with Position-Enhanced Graph Neural Network". In *WWW* 2020.

[17] David Jurgens and Mohammad Taher Pilehvar. "SemEval-2016 Task 14: Semantic Taxonomy Enrichment". In *NAACL* 2016.

[18] Nikhita Vedula, Patrick K. Nicholson, Deepak Ajwani, Sourav Dutta, A. Sala, and S. Parthasarathy. "Enriching Taxonomies With Functional Domain Knowledge". In *SIGIR* 2018.

[19] Yu, Yue, Yinghao Li, Jiaming Shen, Haoyang Feng, Jimeng Sun, and Chao Zhang. "STEAM: Self-Supervised Taxonomy Expansion with Mini-Paths". In *KDD* 2020.

[20] Emaad Manzoor, Rui Li, Dhananjay Shrouty, and Jure Leskovec. "Expanding Taxonomies with Implicit Edge Semantics". In *WWW* 2020.

[21] Qingkai Zeng, Jinfeng Lin, Wenhao Yu, J. Cleland-Huang, and Meng Jiang. "Enhancing Taxonomy Completion with Concept Generation via Fusing Relational Representations". In *KDD* 2021.

[22] Jieyu Zhang, Xiangchen Song, Ying Zeng, Jiaze Chen, Jiaming Shen, Yuning Mao, and Lei Li. "Taxonomy Completion via Triplet Matching Network". In *AAAI* 2021.

[23] Yangqiu Song and Dan Roth. "On Dataless Hierarchical Text Classification." In *AAAI* 2014.

[24] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. "Weakly-Supervised Hierarchical Text Classification". In *AAAI* 2019.

[25] Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. "Text Classification Using Label Names Only: A Language Model Self-Training Approach". In *EMNLP* 2020.

[26] Zihan Wang, Dheeraj Mekala, and Jingbo Shang. "X-Class: Text Classification with Extremely Weak Supervision". In *NAACL* 2021.

[27] Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren, and Jiawei Han. "TaxoClass: Hierarchical Multi-Label Text Classification Using Only Class Names". In *NAACL* 2021.

[28] Jin Huang, Zhaochun Ren, Wayne Xin Zhao, Gaole He, Ji-Rong Wen, and Daxiang Dong. "Taxonomy-Aware Multi-Hop Reasoning Networks for Sequential Recommendation". In *WSDM* 2019.

[29] Bang Liu, Weidong Guo, Di Niu, Chaoyue Wang, Shunnan Xu, Jinghong Lin, Kunfeng Lai, and Yu Xu. "A User-Centered Concept Mining System for Query and Document Understanding at Tencent". In *KDD* 2019.