Constructing and Mining Heterogeneous Information Networks from Massive Text

Jingbo Shang, Jiaming Shen, Liyuan Liu, Jiawei Han
Department of Computer Science, University of Illinois Urbana-Champaign, IL, USA
{shang7, js2, ll2, hanj}@illinois.edu

ABSTRACT

Real-world data exists largely in the form of unstructured texts. A grand challenge on data mining research is to develop effective and scalable methods that may transform *unstructured text* into *structured knowledge*. Based on our vision, it is highly beneficial to transform such text into *structured heterogeneous information networks*, on which *actionable knowledge* can be generated based on the user's need. In this tutorial, we provide a comprehensive overview on recent research and development in this direction. First, we introduce a series of effective methods that construct *heterogeneous information networks* from massive, domain-specific text corpora. Then we discuss methods that mine such text-rich networks based on the user's need. Specifically, we focus on scalable, effective, weakly supervised, language-agnostic methods that work on various kinds of text. We further demonstrate, on real datasets (including news articles, scientific publications, and product reviews), how information networks can be constructed and how they can assist further exploratory analysis.

KEYWORDS

Phrase Mining; Entity Recognition; Taxonomy Construction; Network Mining and Applications; Massive Text Corpora

ACM Reference Format:
Jingbo Shang, Jiaming Shen, Liyuan Liu, Jiawei Han. 2019. Constructing and Mining Heterogeneous Information Networks from Massive Text. In *Proceedings of The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3292500.3332275

WHAT WILL BE COVERED IN THIS TUTORIAL?

This tutorial presents a comprehensive overview of recent developments on mining structures from text and multidimensional text analysis (see major technical contents and detailed outline). We will discuss the following key issues: (1) phrase mining from massive, domain-specific text corpora; (2) named entity recognition using contextualized representation learning; (3) taxonomy construction from raw texts with little supervision; and (4) the impact and applications of the constructed text-rich network in knowledge discovery.

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08.
https://doi.org/10.1145/3292500.3332275

WHY IS THIS TUTORIAL IMPORTANT?

Mining structures and knowledge from unstructured text is one of the major challenges in data mining with huge potential impacts since the majority of the real-world big data are unstructured and are in various forms of text. So far, NLP, machine learning, or text mining communities have not paid enough attention to exploring automated construction of structured, typed information networks. At the same time, the data mining community has demonstrated the tremendous power of structured analysis using typed information networks. Unfortunately, there are not so many existing structured information networks, and it is nontrivial to turn massive unstructured text data into structures and construct such typed structures. This tutorial is to bridge this gap and explore the power of data-driven construction and analysis of massive heterogeneous information network.

In this tutorial, we overview recent developments in this direction, with a focus on effort-light approach, which puts less burden on humans to annotate such data (hence more scalable) but also achieves high quality on structure discovery. We will cover four major themes: (1) *phrase mining* for both entity and relational phrases, (2) *named entity recognition* for typed nodes, (3) *taxonomy construction* for better network exploration, and (4) *structured analysis* on constructed networks.

DETAILED TUTORIAL OUTLINE

We first present a data-drive approach to construct structured networks using phrase mining and entity recognition. We then discuss how to better explore the constructed networks using topic taxonomies and network mining methods. We put more emphasis on recent state-of-the-art, automated methods. The term "automated" here means using weak (i.e., small sets of annotated data) or distant supervision using general, public knowledge base (e.g., Wikipedia).

One important feature of this tutorial is that we interleave the principle and method introduction with the system demos to show how some newly developed methods work on various kinds of real-world data sets effectively and efficiently. We will introduce the related, open-source software packages as well. The detailed outline of the topics that will be covered in the tutorial can be found in our website: https://shangjingbo1226.github.io/2019-04-22-kdd-tutorial/. A brief version is as follows.

- Introduction: Motation & Overview
- Phrase Mining [2, 5, 10, 11, 19]
- Named Entity Recognition [3, 6, 9, 12, 15, 18, 20]
- Taxonomy Construction [1, 7, 13, 16, 17, 23–25, 29, 30]
- Mining Constructed Networks [4, 8, 14, 21, 26–28]
- System Demos [22, 27]
- Summary and Future Directions
- Question Answering and Discussions

TUTORS AND PAST TUTORIAL EXPERIENCES

We have four tutors. All are contributors to the tutorial. Jiawei Han and at least two others will be presenting (based on the acceptance of their other KDD submissions).

- **Jingbo Shang**, Ph.D. candidate, Computer Science, Univ. of Illinois at Urbana-Champaign. His research focuses on mining and constructing structured knowledge from massive text corpora with minimum human effort. His research has been recognized by multiple prestigious awards, including Grand Prize of Yelp Dataset Challenge (2015), Google PhD Fellowships (2017–2019) on Structured Data and Database Management. Mr. Shang has rich experiences in delivering tutorials in major conferences (SIGMOD'17, WWW'17, SIGKDD'17, and SIGKDD'18).
- **Jiaming Shen**, Ph.D. candidate, Computer Science, UIUC. His research focuses on taxonomy construction from massive text corpora, and taxonomy-aware semantic search and ranking. He is the recipient of Brian Totty Graduate Fellowship in 2016.
- **Liyuan Liu**, Ph.D. candidate, Computer Science, UIUC. His research interest mainly lies in data-driven text mining, including contextualized representations with language modeling, weak and heterogeneous supervision.
- **Jiawei Han**, Abel Bliss 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 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.*, WSDM 2018 keynote).

ACKNOWLEDGMENTS

Research was sponsored in part by U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), DARPA under Agreement No. W911NF-17-C-0099, National Science Foundation IIS 16-18481, IIS 17-04532, and IIS-17-41317, DTRA HDTRA11810026, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov).

REFERENCES

[1] David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. 2003. Hierarchical Topic Models and the Nested Chinese Restaurant Process. In *NIPS*.

[2] Marina Danilevsky, Chi Wang, Nihit Desai, Jingyi Guo, and Jiawei Han. 2013. Kert: Automatic extraction and ranking of topical keyphrases from content-representative document titles. *arXiv preprint arXiv:1306.0271* (2013).

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[4] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 135–144.

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

[6] Jason Fries, Sen Wu, Alex Ratner, and Christopher Ré. 2017. SwellShark: A Generative Model for Biomedical Named Entity Recognition without Labeled Data. *arXiv preprint arXiv:1704.06360* (2017).

[7] Ruiji Fu, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning Semantic Hierarchies via Word Embeddings. In *ACL*.

[8] Rana Hussein, Dingqi Yang, and Philippe Cudré-Mauroux. 2018. Are Meta-Paths Necessary?: Revisiting Heterogeneous Graph Embeddings. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. ACM, 437–446.

[9] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360* (2016).

[10] Bing Li, Xiaochun Yang, Bin Wang, and Wei Cui. 2017. Efficiently Mining High Quality Phrases from Texts.. In *AAAI*. 3474–3481.

[11] Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han. 2015. Mining Quality Phrases from Massive Text Corpora. In *SIGMOD*.

[12] Liyuan Liu, Jingbo Shang, Frank Xu, Xiang Ren, Huan Gui, Jian Peng, and Jiawei Han. 2017. Empower Sequence Labeling with Task-Aware Neural Language Model. *arXiv preprint arXiv:1709.04109* (2017).

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

[14] Yike Liu, Tara Safavi, Abhilash Dighe, and Danai Koutra. 2018. Graph summarization methods and applications: A survey. *ACM Computing Surveys (CSUR)* 51, 3 (2018), 62.

[15] Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354* (2016).

[16] Yuning Mao, Xiang Ren, Jiaming Shen, Xiaotao Gu, and Jiawei Han. 2018. End-to-End Reinforcement Learning for Automatic Taxonomy Induction. In *ACL*.

[17] David M. Mimno, Wei Li, and Andrew D McCallum. 2007. Mixtures of hierarchical topics with Pachinko allocation. In *ICML*.

[18] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Vol. 1. 2227–2237.

[19] 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* 30, 10 (2018), 1825–1837.

[20] Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, 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*. 2054–2064.

[21] Jingbo Shang, Meng Qu, Jialu Liu, Lance M Kaplan, Jiawei Han, and Jian Peng. 2016. Meta-Path Guided Embedding for Similarity Search in Large-Scale Heterogeneous Information Networks. *arXiv preprint arXiv:1610.09769* (2016).

[22] Jingbo Shang, Qi Zhu, Jiaming Shen, Xuan Wang, Xiaotao Gu, Lance Kaplan, Timothy Harratty, and Jiawei Han. 2018. AutoNet: Automated Network Construction and Exploration System from Domain-Specific Corpora. (2018).

[23] Jiaming Shen, Zeqiu Wu, Dongming Lei, Jingbo Shang, Xiang Ren, and Jiawei Han. 2017. Setexpan: Corpus-based set expansion via context feature selection and rank ensemble. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 288–304.

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

[25] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic Taxonomy Induction from Heterogenous Evidence. In *ACL*.

[26] Fangbo Tao, Chao Zhang, Xiusi Chen, Meng Jiang, Tim Hanratty, Lance Kaplan, and Jiawei Han. 2018. Doc2Cube: Allocating Documents to Text Cube without Labeled Data. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 1260–1265.

[27] Fangbo Tao, Honglei Zhuang, Chi Wang Yu, Qi Wang, Taylor Cassidy, Lance R. Kaplan, Clare R. Voss, and Jiawei Han. 2016. Multi-Dimensional, Phrase-Based Summarization in Text Cubes. *IEEE Data Eng. Bull.* 39, 3 (2016), 74–84. http://sites.computer.org/debull/A16sept/p74.pdf

[28] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex Embeddings for Simple Link Prediction. In *ICML*.

[29] Chi Wang, Marina Danilevsky, Nihit Desai, Yinan Zhang, Phuong Nguyen, Thrivikrama Taula, and Jiawei Han. 2013. A phrase mining framework for recursive construction of a topical hierarchy. In *KDD*.

[30] Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian M. Sadler, Michelle T. Vanni, and Jiawei Han. 2018. TaxoGen: Constructing Topical Concept Taxonomy by Adaptive Term Embedding and Clustering. In *KDD*.