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.

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 Research was sponsored in part by U.S. Army Research Lab. under.

REFERENCES


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


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


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


[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.


[22] Jingbo Shang, Qi Zhu, Jiawei 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).


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

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

[26] Fangbo Tao, Chao Zhang, Xiaowu Chen, Meng Jiang, Tim Harratty, Lance Kaplan, and Jiawei Han. 2018. Docs2Cube: Allocating Documents to Text Cube without Labeled Data. In In 2018 IEEE International Conference on Data Mining (ICDM). IEEE, 1260–1265.


[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, Xiaowu 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.

Jingbo Shang, Jiaming Shen, Liyuan Liu, Jiawei Han