From Unstructured Text to TextCube: Automated Construction and Multidimensional Exploration

Jiawei Han
University of Illinois, Urbana Champaign, 201 N Goodwin Ave, Urbana, IL 61801, USA
hanj@illinois.edu

ABSTRACT

The real-world big data are largely unstructured, interconnected, and dynamic, in the form of natural language text. It is highly desirable to transform such massive unstructured data into structured knowledge. Many researchers rely on labor-intensive labeling and curation to extract knowledge from such data, which may not be scalable, especially considering that a lot of text corpora are highly dynamic and domain specific. We believe that massive text data itself may disclose a large body of hidden patterns, structures, and knowledge. With domain-independent and domain-dependent knowledge bases, we propose to explore the power of massive data itself for turning unstructured data into structured knowledge. By organizing massive text documents into multidimensional text cubes, we show structured knowledge can be extracted and used effectively. In this talk, we introduce a set of methods developed recently in our group for such an exploration, including mining quality phrases, entity recognition and typing, multi-faceted taxonomy construction, and construction and exploration of multi-dimensional text cubes. We show that data-driven approach could be a promising direction at transforming massive text data into structured knowledge.

CCS CONCEPTS

• **Information systems** → *Data mining*; • **Applied computing** → **Enterprise ontologies, taxonomies and vocabularies**.

KEYWORDS

Data mining, text mining, text embedding, textcube construction

ACM Reference Format:
Jiawei Han. 2019. From Unstructured Text to TextCube: Automated Construction and Multidimensional Exploration. In *The 28th ACM International Conference on Information and Knowledge Management (CIKM '19), November 3–7, 2019, Beijing, China.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3357384.3358172

INTRODUCTION

Motivated by developing automated methods for transforming massive unstructured text data into structured knowledge, we have proposed a multidimensional text cube model and developed a set of methods for automatically allocating unstructured text documents in multidimensional text cube structures, which may strike a key step for turning unstructured text into structured knowledge. Besides presenting our vision, we will introduce a set of concrete methods developed recently in our group towards such an exploration, including mining quality phrases [3], spherical text embedding [1], entity recognition and typing [6], multi-faceted taxonomy construction [4, 8], and construction and exploration of multi-dimensional text cubes [2, 5, 7]. We show that data-driven approach could be a promising direction at transforming massive text data into structured knowledge.

CIKM '19, November 3–7, 2019, Beijing, China

ACM ISBN 978-1-4503-6976-3/19/11.
https://doi.org/10.1145/3357384.3358172

SHORT BIOGRAPHY

Jiawei Han is Michael Aiken Chair Professor in the Department of Computer Science, University of Illinois at Urbana-Champaign. He has been researching into data mining, information network analysis, database systems, and data warehousing, with over 900 journal and conference publications. He has chaired or served on many program committees of international conferences in most data mining and database conferences. He also served as the founding Editor-In-Chief of ACM Transactions on Knowledge Discovery from Data, the Director of Information Network Academic Research Center supported by U.S. Army Research Lab (2009-2016), and the co-Director of KnowEnG, an NIH funded Center of Excellence in Big Data Computing (2014-2019) He is Fellow of ACM, Fellow of IEEE, and received 2004 ACM SIGKDD Innovations Award, 2005 IEEE Computer Society Technical Achievement Award, 2009 M. Wallace McDowell Award from IEEE Computer Society, and 2018 Japan's Funai Achievement Award. His co-authored book "Data Mining: Concepts and Techniques" has been adopted as a textbook popularly worldwide.

ACKNOWLEDGEMENTS

Research was sponsored in part by U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), DARPA under Agreements No. W911NF-17-C-0099 and No. FA8750-19-2-1004, 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] Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance Kaplan, and Jiawei Han. Dec. 2019. Spherical Text Embedding. In *Proc. 2019 Conf. on Neural Information Processing Systems (NeurIPS'19).*
[2] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2019. Weakly-Supervised Hierarchical Text Classification. In *Proc. 2019 AAAI Conf. on Artificial Intelligence, (AAAI'19).*

[3] Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R. Voss, and Jiawei Han. 2018. Automated Phrase Mining from Massive Text Corpora. *IEEE Trans. Knowl. Data Eng.* 30, 10 (2018), 1825–1837.

[4] 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 *Proc. 2018 ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining (KDD'18).*

[5] Fangbo Tao, Chao Zhang, Xiusi Chen, Meng Jiang, Tim Hanratty, Lance M. Kaplan, and Jiawei Han. Doc2Cube: Allocating Documents to Text Cube Without Labeled Data. In *Proc. of 2018 IEEE Int. Conf. on Data Mining (ICDM'18).*

[6] Xuan Wang, Yu Zhang, Qi Li, Xiang Ren, Jingbo Shang, and Jiawei Han. 2019. Supervised Biomedical Named Entity Recognition with Dictionary Expansion. In *Proc. of 2019 IEEE Int. Conf. on Bioinformatics and Biomedicine (IEEE-BIBM'19).*

[7] Chao Zhang and Jiawei Han. 2019. *Multidimensional Mining of Massive Text Data.* Morgan & Claypool Publishers.

[8] Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian M. Sadler, Michelle Vanni, and Jiawei Han. TaxoGen: Unsupervised Topic Taxonomy Construction by Adaptive Term Embedding and Clustering. In *Proc. 2018 ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining (KDD'18).*